

# *De la IA Generativa a la IA Responsable*

*Javier Vázquez-Salceda*

*Cap del grup de recerca KEMLG-IDEAI*

*Vicedegà de IA del Col·legi d'Enginyeria en Informàtica de Catalunya*

*Membre del Grup Intercol·legial de Treball en IA (GTAI)*

*[jvazquez@cs.upc.edu](mailto:jvazquez@cs.upc.edu)*

*<http://www.cs.upc.edu/~jvazquez>*



Knowledge Engineering and Machine Learning Group  
UNIVERSITAT POLITÈCNICA DE CATALUNYA

*<https://kemlg.upc.edu>*

informàTICa  
Enginyeria en informàtica  
de Catalunya

# No tota la IA es IA Generativa

---



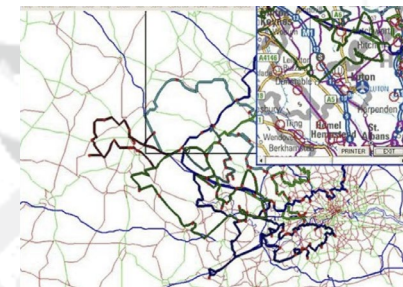
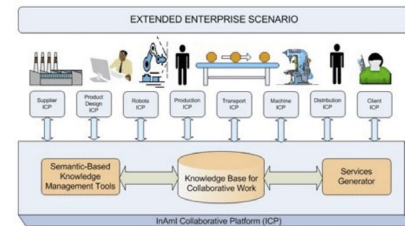
Knowledge Engineering and Machine Learning Group  
UNIVERSITAT POLITÈCNICA DE CATALUNYA

<https://kemlg.upc.edu>

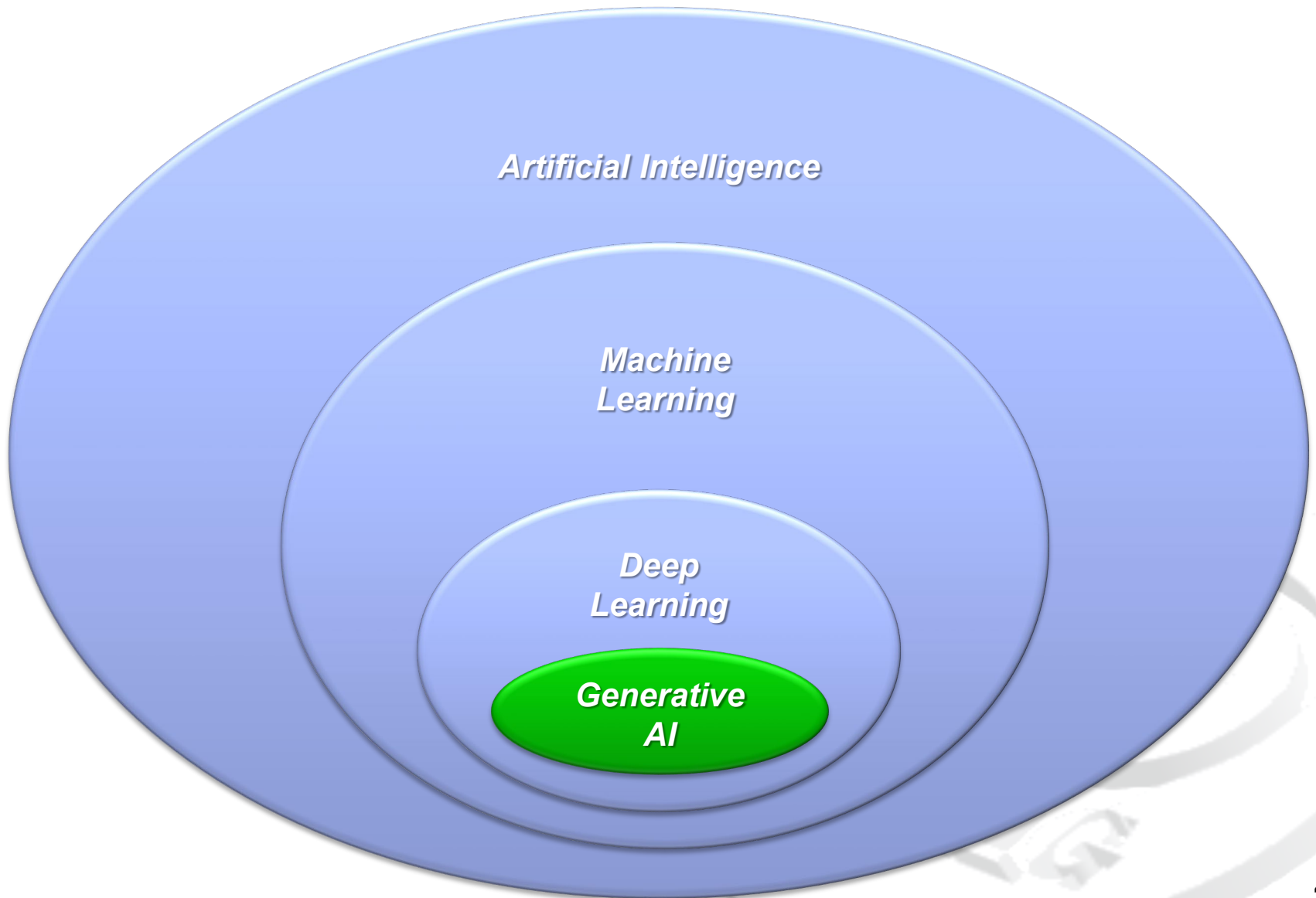
informàTICa  
Enginyeria en informàtica  
de Catalunya

# La IA porta molt temps amb nosaltres

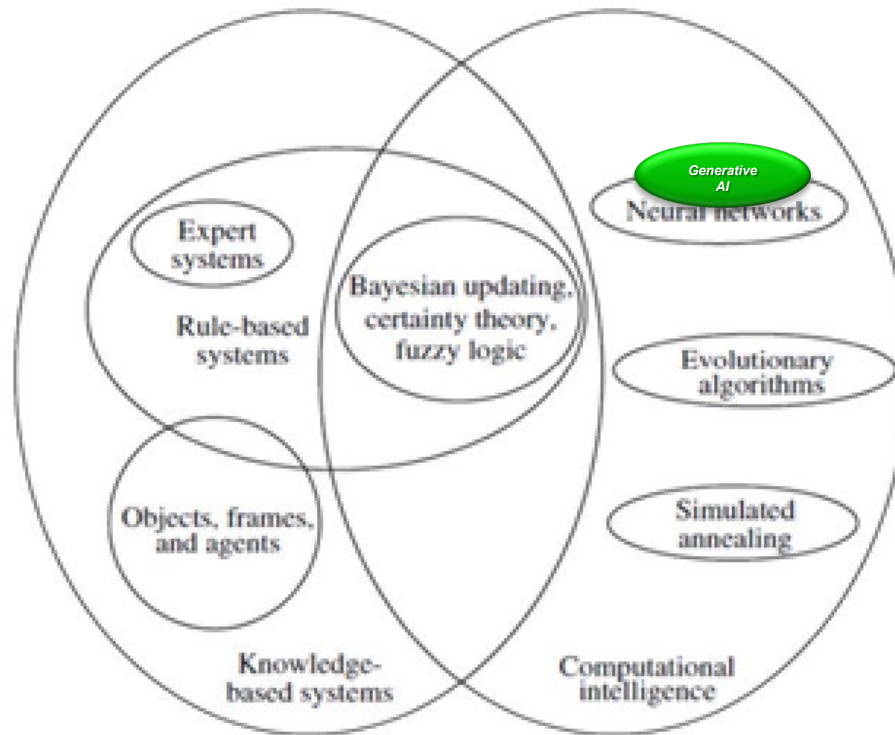
- Els orígens de la Intel·ligència Artificial es remunten als filòsofs de la antiga Grècia (*com mecanitzar el pensament humà?*)
- Es una àrea de la ciència pròpia des de 1956
- Te influències molt diverses
  - *Filosofia, Matemàtica, Ciències de la Computació, Psicologia, Biologia, ...*
- Te molts usos beneficiosos i no perillosos
  - *Medicina*
  - *Gestió mediambiental*
  - *Pedagogia*
  - *Manufactura intel·ligent*
  - *Detecció i predicció de fallides*
  - *Logística...*



Hi ha molta IA més enllà de la IA Generativa...

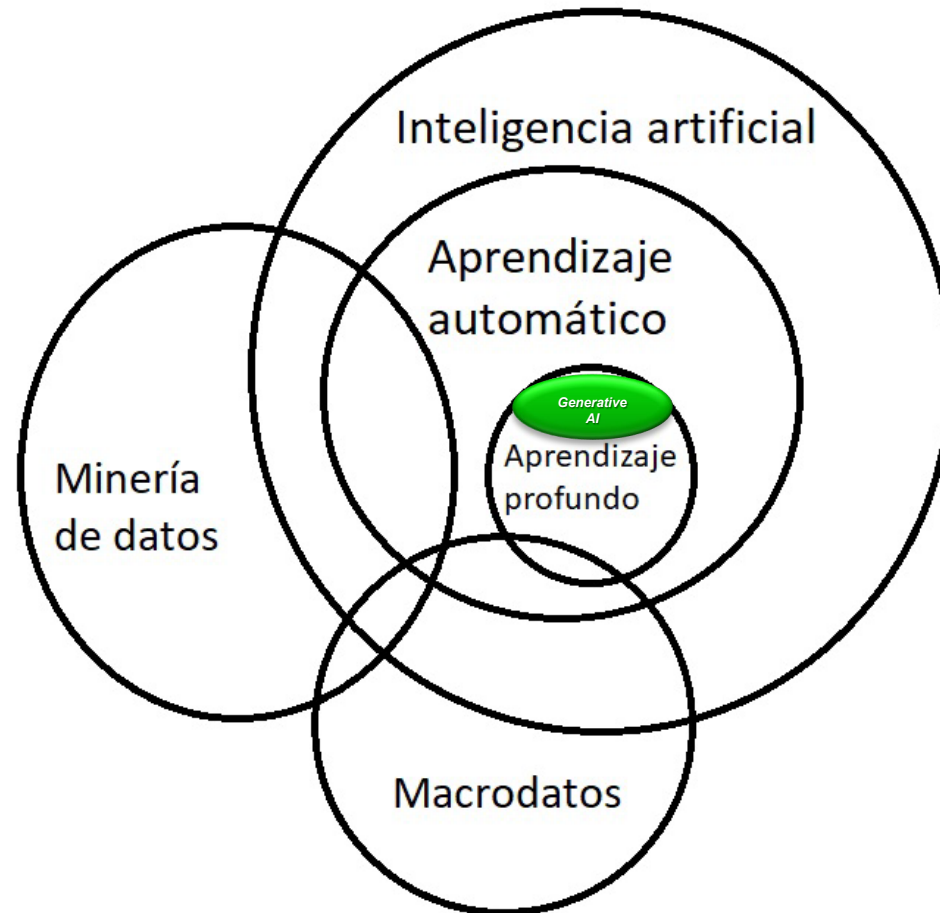


# Hi ha molta IA més enllà de la IA Generativa...



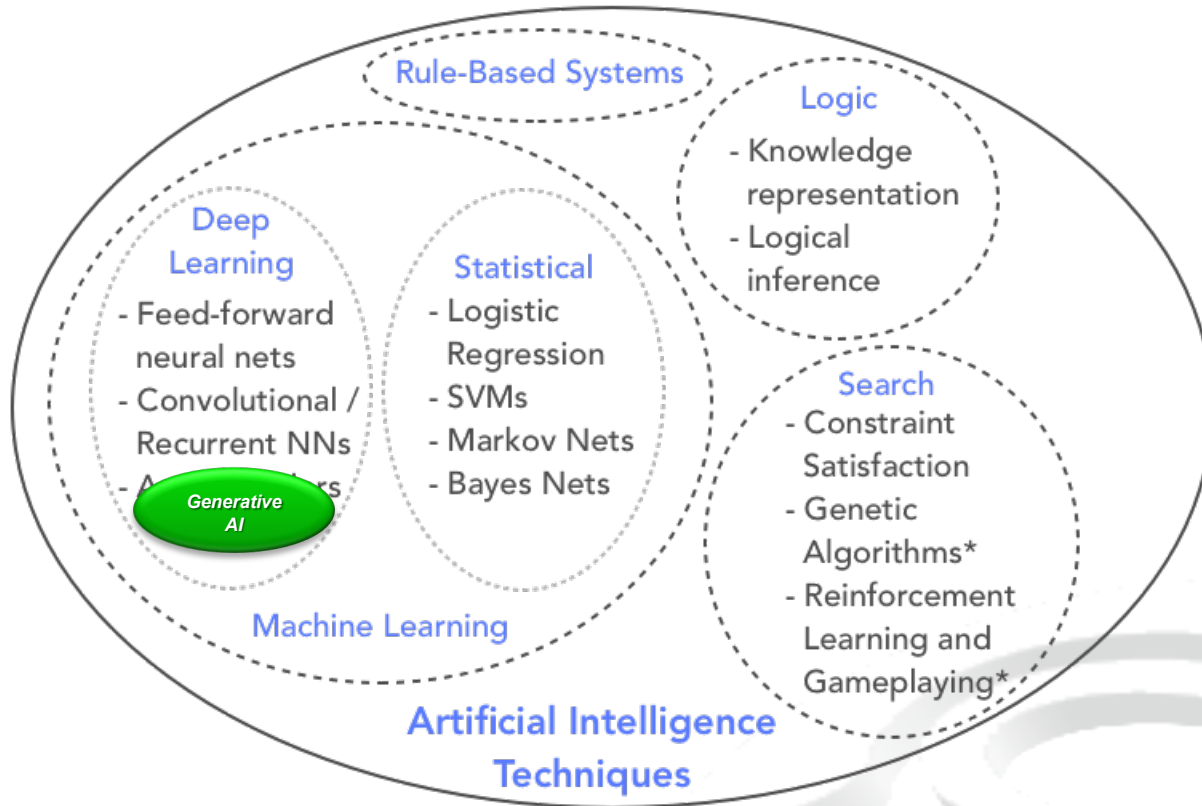
[A. Hopgood (2001)]

# Hi ha molta IA més enllà de la IA Generativa...



[Salehi i Burgeño (2018)]

# Hi ha molta IA més enllà de la IA Generativa...



[S. Shroufi (2018)]

# Perquè es parla tant (ara) d'IA (generativa)?

- La capacitat de la IA de **resoldre problemes cada cop més complexos** esta lligada a la capacitat de computació

50's – 80's	90's - 2014	2015 -
<ul style="list-style-type: none"> <li>• Velocitat de processament (CPU)</li> <li>• Capacitat de memòria limitada (1 ordinador)</li> </ul>	<ul style="list-style-type: none"> <li>• Velocitat de processament (CPU) en Mhz-Ghz</li> <li>• Ordinadors en xarxa, memòria distribuïda</li> <li>• Connexions de banda ampla</li> </ul>	<ul style="list-style-type: none"> <li>• Velocitat de processament (CPU i GPU) en Ghz</li> <li>• Ordinadors en xarxa, memòria distribuïda</li> <li>• Connexions d'alta velocitat</li> <li>• Grans volums de dades</li> </ul>

- Recentment s'han posat a l'abast nombroses eines de **Deep Learning** i **IA Generativa** que poden ser usades per gent sense una (amplia) formació en Intel·ligència Artificial
  - *“Dona'm moltes dades per entrenar i et generaré un model”*
  - Això es una **oportunitat per la innovació** però també **un perill** en mans no expertes



# Tipus i usos de la IA Generativa

---



Knowledge Engineering and Machine Learning Group  
UNIVERSITAT POLITÈCNICA DE CATALUNYA

<https://kemlg.upc.edu>




informàTICa  
Enginyeria en informàtica  
de Catalunya

# Tipus i usos de la IA Generativa





- **Generació de text**
  - resums, documents, programes d'ordinador...
- **Generació de so**
  - Veu parlada, música, veu cantada, efectes sonors...
- **Generació d'imatge (2D, 3D) i vídeo**
  - Disseny gràfic, fotografia, cinematografia, arts plàstiques...
- **Altres Generacions**
  - Generació d'accions, generació de comportament, generació (predictiva) de mapes del temps...

# Usos de la IA Generativa: text




## ■ Traducció automàtica

- Traducció de textos amb una estructura menys rígida, més natural 
- Traducció de grans volums de text en múltiples idiomes 
- Traducció en certs dominis (medicina, legal): encara te perill de que la traducció no sigui correcta, requereix supervisió d'un humà 

## ■ Autocompletat / Correcció de text

- Correcció de text a nivell de frase (millora la qualitat del text)  
- Autocompletat (predicció) del text (millora de la productivitat) 
- Perill: l'usuari es deixa de preocupar per saber com escriure bé 

## ■ Descripció d'imatges (Image-to-text)

- Descriure els elements d'una escena/d'un video 
- Accesibilitat d'invidents a continguts visuals  

# Usos de la IA Generativa: text

## ■ Creació de resums de documents

- Resums d'una base de documents per facilitar la cerca
- Resums de missatges de correu (millora de la productivitat)
- Perill: Resums entregats com a treballs fets per l'alumne



## ■ Creació de documents sencers

- Generar redactats complets de parts molt repetitives (e.g., plecs de condicions)
- Generar redactats complets a partir d'un esquema avançat del document i el seu contingut
- Generar documents sencers a partir d'un prompt
- Perill: *Fake documents*
- Perill: Documents entregats com a treballs fets per l'alumne



# Usos de la IA Generativa: text

## ■ Creació de codi de programa

- Autocompletat de codi (millora de la productivitat)
- Correcció de codi (millora de la qualitat)
- Correcció de codi (en l'aprenentatge de la programació)
- Generació automàtica de codi sencer a partir d'un prompt en el cas d'un programador expert
- Generació automàtica de codi sencer a partir d'un prompt en absència d'un programador expert



**COMPTE!**



Avui dia hi ha un munt de sistemes controlats per programes d'ordinador. Un tros de codi generat automàticament que sigui incorrecte pot tenir greus conseqüències (aturades, accidents)

***Qui es responsable de les conseqüències? La IA?***

*L'humà que ha utilitzat la IA?*

# Usos de la IA Generativa: text

## ■ Chatbots

- Interacció escrita de forma més natural
- Facilita la interacció entre persones i sistemes
- Risc: part del text generat no es correcte
- Perill: no s'informa a l'usuari de que la resposta pot contenir errades o imprecisions



**COMPTE!**



El que un chatbot generi respostes potencialment incorrectes o imprecises pot generar riscos directament a la persona que l'usa, o a terceres persones.

Veurem un parell d'exemples amb el ChatGPT

# Generació de text: Risc directe

## ChatGPT com a cercador d'informació

- Suposem un resident legal d'un altre país que ha de renovar el permís de residència
- No té formació en lleis ni en cerca d'informació a Internet
- ChatGPT es fa fàcil d'usar i dona respostes més entenedores, s'ha acostumat a fer-lo servir com a cercador
- Pregunta en Espanyol:  
"como renovar permiso de residencia en Barcelona"
- ChatGPT utilitza TOTES les pàgines web en Espanyol sobre renovació dels permisos de residència per "crear" el text de resposta.
- Les dates i alguns passos són incorrectes. Però ChatGPT no avisa clarament que és una resposta creada artificialment, i que poden haver errors
- **L'usuari perd la residència i es veu obligat a retornar al país d'origen (amb la seva família)**



# Generació de text: Risc indirecte

## ChatGPT ens explica com fer Aprenentatge Automàtic

MI Dispongo de un listado de proyectos de todo tipo. Hay proyectos que son interesantes y otros que no son interesantes. Necesito un algoritmo que aprenda a decirme los proyectos que son interesantes para mi a través de los proyectos que voy marcando como interesantes.



Lo que estás buscando es un algoritmo de aprendizaje automático supervisado que pueda aprender a clasificar los proyectos como "interesantes" o "no interesantes" en función de los proyectos que hayas marcado previamente como interesantes. Para ello, puedes seguir los siguientes pasos:



1. Recopila un conjunto de datos que contenga información sobre los proyectos, incluyendo aquellos que has marcado como interesantes y aquellos que no. Esta información puede incluir descripciones del proyecto, el tipo de proyecto, la duración, el presupuesto, el equipo necesario, etc.
2. Crea un modelo de aprendizaje automático supervisado que pueda aprender a clasificar los proyectos como interesantes o no interesantes en función de las características del proyecto. Para ello, puedes utilizar algoritmos como árboles de decisión, regresión logística o redes neuronales.
3. Entrena el modelo utilizando los datos recopilados en el paso 1. El modelo aprenderá a identificar patrones en los datos que permitan clasificar los proyectos como interesantes o no interesantes.
4. Evalúa el modelo utilizando datos de prueba. Para ello, utiliza un conjunto de datos diferente al utilizado para el entrenamiento del modelo. De esta forma, podrás evaluar la precisión del modelo para clasificar proyectos como interesantes o no interesantes.
5. Utiliza el modelo entrenado para clasificar nuevos proyectos como interesantes o no interesantes en función de sus características. De esta forma, podrás tomar decisiones informadas sobre qué proyectos son más adecuados para ti.

Es importante destacar que la precisión del modelo dependerá en gran medida de la calidad y cantidad de los datos utilizados para entrenar el modelo. Por lo tanto, es recomendable recopilar la mayor cantidad posible de datos relevantes sobre los proyectos para obtener resultados más precisos.



# Generació de text: Risc indirecte

## ChatGPT ens explica com fer Aprenentatge Automàtic

MI

Qué plataforma es la mejor para crear un algoritmo de aprendizaje automático?



Hay varias plataformas para crear algoritmos de aprendizaje automático, y la mejor plataforma para ti dependerá de tus necesidades y conocimientos en programación. Algunas de las opciones más populares son:



1. Python: Python es un lenguaje de programación popular para el aprendizaje automático. Tiene numerosas bibliotecas de aprendizaje automático, como Scikit-learn, TensorFlow, Keras, PyTorch y Pandas, que hacen que sea fácil crear modelos de aprendizaje automático.
2. R: R es un lenguaje de programación popular para el análisis estadístico y la visualización de datos, pero también tiene numerosas bibliotecas para el aprendizaje automático, como Caret, MLR y TensorFlow.
3. MATLAB: MATLAB es un entorno de programación numérica popular para la ingeniería y la ciencia. Tiene una amplia gama de funciones y herramientas de aprendizaje automático, como la caja de herramientas de aprendizaje automático de MATLAB y Deep Learning Toolbox.
4. Weka: Weka es una plataforma de código abierto que proporciona una interfaz gráfica para crear y evaluar modelos de aprendizaje automático. Tiene una amplia gama de algoritmos de aprendizaje automático preconstruidos y es adecuada para aquellos que no tienen experiencia en programación.
5. Google Cloud Platform: Google Cloud Platform ofrece una amplia gama de herramientas y servicios para el aprendizaje automático, como TensorFlow, AutoML y BigQuery. Es una opción adecuada para aquellos que buscan una plataforma de nube y tienen experiencia en programación.

En resumen, la mejor plataforma para crear un algoritmo de aprendizaje automático dependerá de tus necesidades y conocimientos en programación. Si eres nuevo en la programación, Weka podría ser una buena opción, mientras que si estás familiarizado con Python, Scikit-learn y TensorFlow podrían ser buenas opciones.

# Generació de text: Risc indirecte

## ChatGPT ens explica com fer Aprenentatge Automàtic

MI

Cómo crear un algoritmo de aprendizaje automático con python?



Para crear un algoritmo de aprendizaje automático con Python, puedes seguir los siguientes pasos:



1. Preparación de datos: El primer paso es recopilar y preparar los datos que se utilizarán para entrenar el modelo. Esto puede incluir la limpieza de datos, la normalización y la transformación de características.
2. Importar bibliotecas de Python: Importa las bibliotecas de aprendizaje automático de Python necesarias, como Numpy, Pandas y Scikit-learn.
3. División del conjunto de datos: Divide los datos en un conjunto de entrenamiento y un conjunto de prueba utilizando la función "train\_test\_split" de Scikit-learn.
4. Selección de un modelo de aprendizaje automático: Hay una variedad de algoritmos de aprendizaje automático que puedes utilizar para clasificación, como regresión logística, árboles de decisión, k-vecinos más cercanos (K-NN), máquinas de soporte vectorial (SVM), redes neuronales, entre otros. Selecciona el modelo que mejor se adapte a tus necesidades.
5. Entrenamiento del modelo: Utiliza la función "fit" de Scikit-learn para entrenar el modelo con el conjunto de entrenamiento.
6. Evaluación del modelo: Utiliza el conjunto de prueba para evaluar la precisión del modelo en la clasificación de los proyectos como interesantes o no interesantes. Utiliza las métricas de evaluación, como la precisión, la sensibilidad, la especificidad y el área bajo la curva (AUC), para evaluar el rendimiento del modelo.
7. Mejora del modelo: Si el modelo no tiene un rendimiento óptimo, se pueden realizar ajustes para mejorar su precisión. Esto puede incluir la selección de características más relevantes, el ajuste de parámetros del modelo y la utilización de técnicas de validación cruzada.
8. Utilización del modelo: Una vez que estés satisfecho con la precisión del modelo, utilízalo para clasificar nuevos proyectos como interesantes o no interesantes en función de sus características.

# Generació de text: Risc indirecte

## ChatGPT ens explica com fer Aprenentatge Automàtic

Aquí tienes un ejemplo básico de cómo entrenar un modelo de regresión logística en Python:

```
python Copy code  
  
# Importar bibliotecas  
import numpy as np  
import pandas as pd  
from sklearn.linear_model import LogisticRegression  
from sklearn.model_selection import train_test_split  
  
# Cargar datos  
data = pd.read_csv("datos_proyectos.csv")  
  
# Dividir los datos en conjunto de entrenamiento y prueba  
X_train, X_test, y_train, y_test = train_test_split(data[['caracteristica_1'  
  
# Crear el modelo de regresión logística  
logreg = LogisticRegression()  
  
# Entrenar el modelo con los datos de entrenamiento  
logreg.fit(X_train, y_train)  
  
# Evaluar el modelo con los datos de prueba  
y_pred = logreg.predict(X_test)  
  
# Calcular la precisión del modelo  
from sklearn.metrics import accuracy_score  
accuracy = accuracy_score(y_test, y_pred)  
print("Precisión del modelo:", accuracy)
```

Recuerda que este es solo un ejemplo básico y que el rendimiento del modelo dependerá de la calidad de los datos utilizados y de la selección del modelo adecuado.

# Generació de text: Risc indirecte

## ChatGPT ens explica com fer Aprenentatge Automàtic

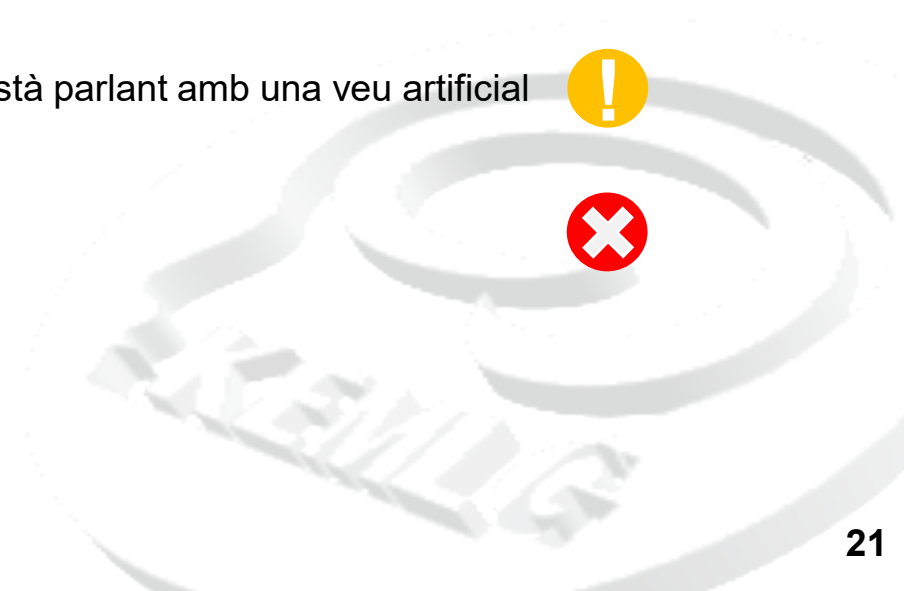
- ***Això es bó o dolent?***
- Depen de quin sigui l'us d'aquesta informació
  - Un expert en IA que está jugant amb el prototip per veure com va, però que no aplicarà el resultat en cap producte.
  - Un alumne (secundària/batxillerat/grau) que te curiositat sobre el Aprenentatge Automàtic i vol fer petits experiments, sense usuaris finals
  - Un desenvolupador no expert en IA que aplicarà les instruccions (i potser el codi!) en un producte amb usuaris finals



# Usos de la IA Generativa: so

## ■ Generació de veu

- Traducció automàtica en temps real (quan es més important la immediatesa que la perfecció)
- Interacció parlada entre Chatbots/Robots i humans
- Accessibilitat per a persones invidents
  - Per exemple: descripció dels objectes en una habitació
- Perill: no informar a l'humà que està parlant amb una veu artificial
- Perill: Suplantació d'identitat



# Usos de la IA Generativa: so

## ■ Generació de música

- Crear acompanyaments/arregaments per la melodia d'un autor
- Transportar una musica a altres tonalitats, a altres estils musicals
- Crear peces senceres sense necessitat d'un autor
- Perill: *Fake music*



## ■ Generació de veu cantada

- Crear veus cantades realistes a partir de text i musica d'un autor
- Crear traduccions de veu cantada
- Suplantació de la veu i tècnica d'un cantant amb consentiment
- Suplantació de la veu i tècnica d'un cantant sense consentiment



La industria musical ja te mecanismes de detecció de frau / plagi / copia basades en IA, que donen suport a regulacions establertes (e.g.: més de 8 compassos iguals es plagi).

# Un bon ús de la generació musical amb IA

- La cantant **Maria Arnal** ha creat **Maria CHOIR**, una peça artística i musical feta amb IA
- S'ha entrenat una IA Generativa per crear una reproducció sintètica de la veu de Maria Arnal
- La veu sintètica harmonitza en temps real el que canten les persones visitants de la instal·lació
- L'objectiu es anar formant un cor híbrid humà-digital que evoluciona al llarg del temps







Més informació:







<https://www.cccb.org/ca/activitats/fitxa/maria-choir/243188>

# Usos de la IA Generativa: imatge i video

## ■ Suport al disseny gràfic

- Creació de logotips, icones, formes a partir de conceptes, idees, esborranys 
- Alteracions del dibuix d'un autor a diferents estils i efectes visuals 
- Passar "a net" esborranys fets a mà (en 2D i 3D) 
- Perill: plagi, còpia 







## ■ Suport a la fotografia

- Correcció automàtica d'imatges de l'autor 
- Exploració artística d'efectes visuals a partir de la imatge de l'usuari 
- Generació d'imatges a partir d'un prompt (*Text to image*)  
- Perill: *Fake photos, Fake news* 
- Perill: Suplantació d'identitat, vulneració del dret a la imatge 



# Usos de la IA Generativa: imatge i video

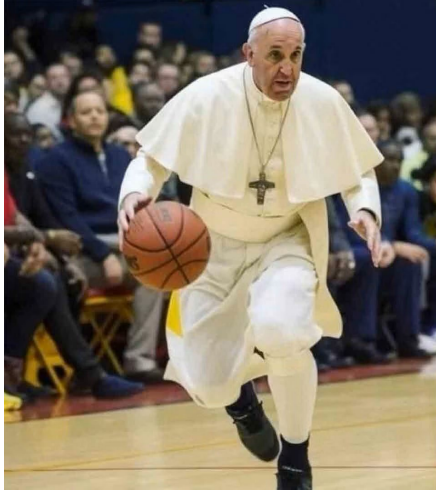
## ■ Suport a la cinematografia

- Correcció automàtica de vídeos 
- Exploració artística d'efectes visuals a partir d'un vídeo de l'usuari 
- Efectes especials realistes 
- Modificació de les faccions de les persones per sincronitzar els llavis a la veu traduïda 
- Generació de videos a partir d'un prompt (*Text to video*) 
- Perill: *Fake vídeos, Fake news* 
- Perill: Suplantació d'identitat, vulneració del dret a la imatge

## ■ Suport a les arts plàstiques

- Generació automàtica d'efectes visuals en temps real a partir de la interacció amb l'artista / el moviment / la música 

# Exemples d'usos (perillosos) de IA Generativa amb imatge o vídeo






*Imatge creada per uns periodistes de **El Mundo** per a reflexionar sobre els perills del Deep Fake*

# Perill extra de la IA Generativa per imatge / vídeo





- Una **IA Generativa per imatge o vídeo** sol tenir com a **objectiu principal** el produir un resultat **el més realista possible**
  - És a dir, que un humà no pugui distingir si es una imatge *real* o *artificial*
- Algunes tècniques d'IA Generativa ho aconseguen **fent competir dos sistemes d'IA**:
  - Un primer sistema de IA que s'entrena per generar imatges/vídeos el més realista possibles
  - Un segon sistema de IA que s'entrena per detectar imatges/vídeos artificials
  - Els resultats que es donen a l'usuari son els que han aconseguit enganyar a la segona IA, entrenada per detectar imatges artificials
    - Sovint es difícil que els humans detectin que la imatge/vídeo és artificial
    - Si ha aconseguit enganyar el segon sistema de IA, es possible que aconsegueixi enganyar altres sistemes de IA de detecció d'imatges/vídeos artificials

# Altres usos de la IA Generativa



## ■ Generació d'accions i plans d'acció

- A partir de la entrada dels sensors, generar les accions que un robot / un cotxe autònom farà a continuació 
- A partir de l'estat actual de la planta, generar les següents comandes als diferents elements de la cadena de producció 
- Perill: accions i plans d'acció basats en una mena d'intuïció digital, no es pot explicar/justificar la acció realitzada o el pla escollit 

## ■ Generació de comportaments socials

- Crear simulacions realistes de models socials per poder analitzar fenòmens  
- Entrenar el comportament social d'un robot en un simulador social, amb humans generats per la IA  

## ■ Generació de models dinàmics

- Generació (predictiva) de mapes de precipitacions 
- Generació (predictiva) de mapes de transit 

# Intel·ligència Artificial Responsable

---



Knowledge Engineering and Machine Learning Group  
UNIVERSITAT POLITÈCNICA DE CATALUNYA

<https://kemlg.upc.edu>

informàTICa  
Enginyeria en informàtica  
de Catalunya

# Futur marc regulatori: EU AI Act

- Objectiu: Assegurar que les IA
  - *ajuden* la gent
  - son una *força per el bé comú* de la societat
- Usos prohibits de la IA
  - Sistemes de *identificació biomètrica* en espais públics en temps real
  - *Categorització de persones* per característiques sensibles (etnia, religió,...)
  - *Sistemes policials predictius* (a partir de perfilat/localització, historial delictiu)
  - *Sistemes de reconeixement d'emocions* en alguns entorns (jutjats, empreses...)
- Usos d'alt risc de la IA
  - Quan pot impactar negativament sobre la salut, seguretat o drets fonamentals de les persones
  - Quan pot afectar/manipular la opinió de les persones
- Obligacions en el cas de les IA de propòsit general
  - Analitzar i mitigar els possibles riscos
    - (salut, seguretat, drets fonamentals, democràcia...)
  - Registrar els models en un registre europeu **abans** de llençar-los al mercat
  - Les IA Generatives han d'avisar que el contingut es generat artificialment

# Guies i principis pel bon ús de la IA

- UNESCO
- Unió Europea
- OECD
- WEF
- Consell d' Europa
- IEEE Ethically Aligned Design
- Estrategies nacionals
- ...

EU HLEG	OECD	IEEE EAD
<ul style="list-style-type: none"> <li>• Human agency and oversight</li> <li>• <b>Technical robustness and safety</b></li> <li>• Privacy and data governance</li> <li>• <b>Transparency</b></li> <li>• <b>Diversity</b>, non-discrimination and fairness</li> <li>• <b>Societal and environmental well-being</b></li> <li>• <b>Accountability</b></li> </ul>	<ul style="list-style-type: none"> <li>• <b>benefit people and the planet</b></li> <li>• respects the rule of law, <b>human rights</b>, democratic values and <b>diversity</b>,</li> <li>• include appropriate safeguards (e.g. human intervention) to ensure a <b>fair and just society</b>.</li> <li>• <b>transparency</b> and responsible disclosure</li> <li>• <b>robust, secure and safe</b></li> <li>• Hold organisations and individuals <b>accountable</b> for proper functioning of AI</li> </ul>	<ul style="list-style-type: none"> <li>• How can we ensure that A/IS do not infringe <b>human rights</b>?</li> <li>• effect of A/IS technologies on <b>human well-being</b>.</li> <li>• How can we assure that designers, manufacturers, owners and operators of A/IS are responsible and <b>accountable</b>?</li> <li>• How can we ensure that A/IS are <b>transparent</b>?</li> <li>• How can we extend the benefits and minimize the risks of AI/AS technology being misused?</li> </ul>

## Idees principals de les guies

- **IA Confiable (Trustworthy AI / TAI)**

fiabilitat basada en 3 components:

- Lícita (complir totes les lleis i els reglaments aplicables)
- Ètica (garanteix el respecte dels principis i valors ètics de la societat)
- Robusta (des del punt de vista tècnic i social, evitar danys accidentals).

- **IA Responsable (Responsible AI / RAI)**

IA desenvolupada amb pràctiques (disseny, desenvolupament i implantació) que

- ajudin positivament a empreses i empleats
- Tinguin un impacte just sobre els clients i la societat

Per saber-ne més:

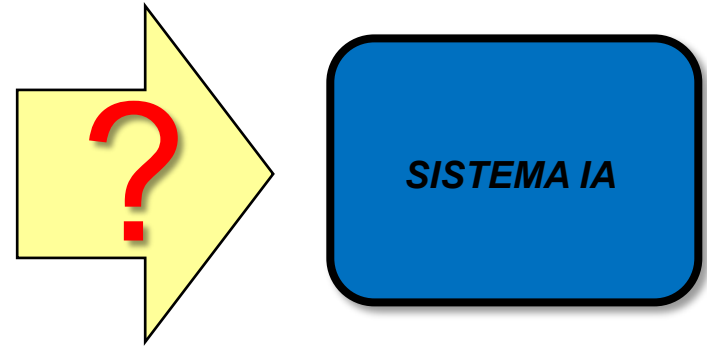
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>



## Com ho apliquem?

- **IA Confiable**

- Lícita?
- Ètica?
- Robusta?



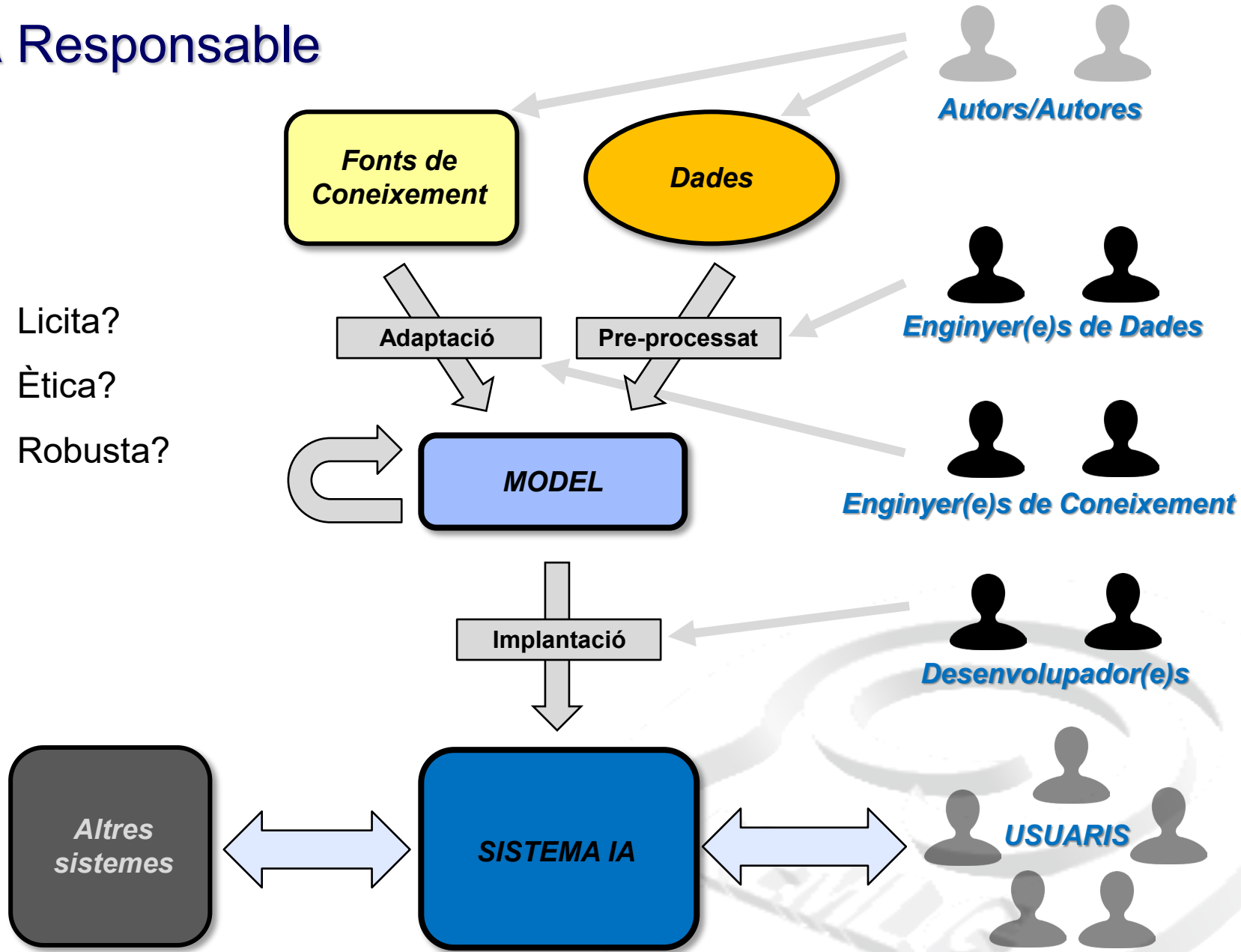
- **IA Responsable**

pràctiques (disseny, desenvolupament i implantació) per

- ajudar positivament (empreses i empleats)
- Tenir un impacte just (clients i societat)

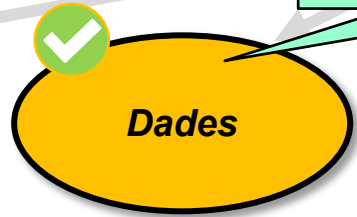
# IA Responsable

- Lícita?
- Ètica?
- Robusta?



# IA Responsable

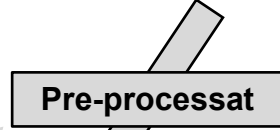
- No vulneren drets
- Obtigudes de forma ètica i legal
- Aliniades amb valors ètics



- No vulneren drets
- Obtigudes de forma ètica i legal
- Aliniades amb valors ètics

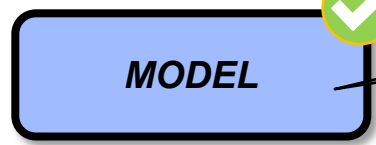
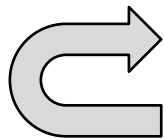
**Autors/Autores**

- Lícita?
- Ètica?
- Robusta?

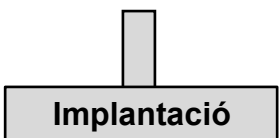


- No vulnera drets
- Promou impactes positius
- Minimització de riscos

**Enginyer(e)s de Dades**



**Enginyer(e)s de Coneixement**



- No vulnera drets
- Resultats ètics i legals
- Anàlisi d'impactes sobre sistema
- Minimització de riscos



- No vulnera drets
- Resultats ètics i legals
- Anàlisi d'impactes sobre usuari
- Minimització de riscos



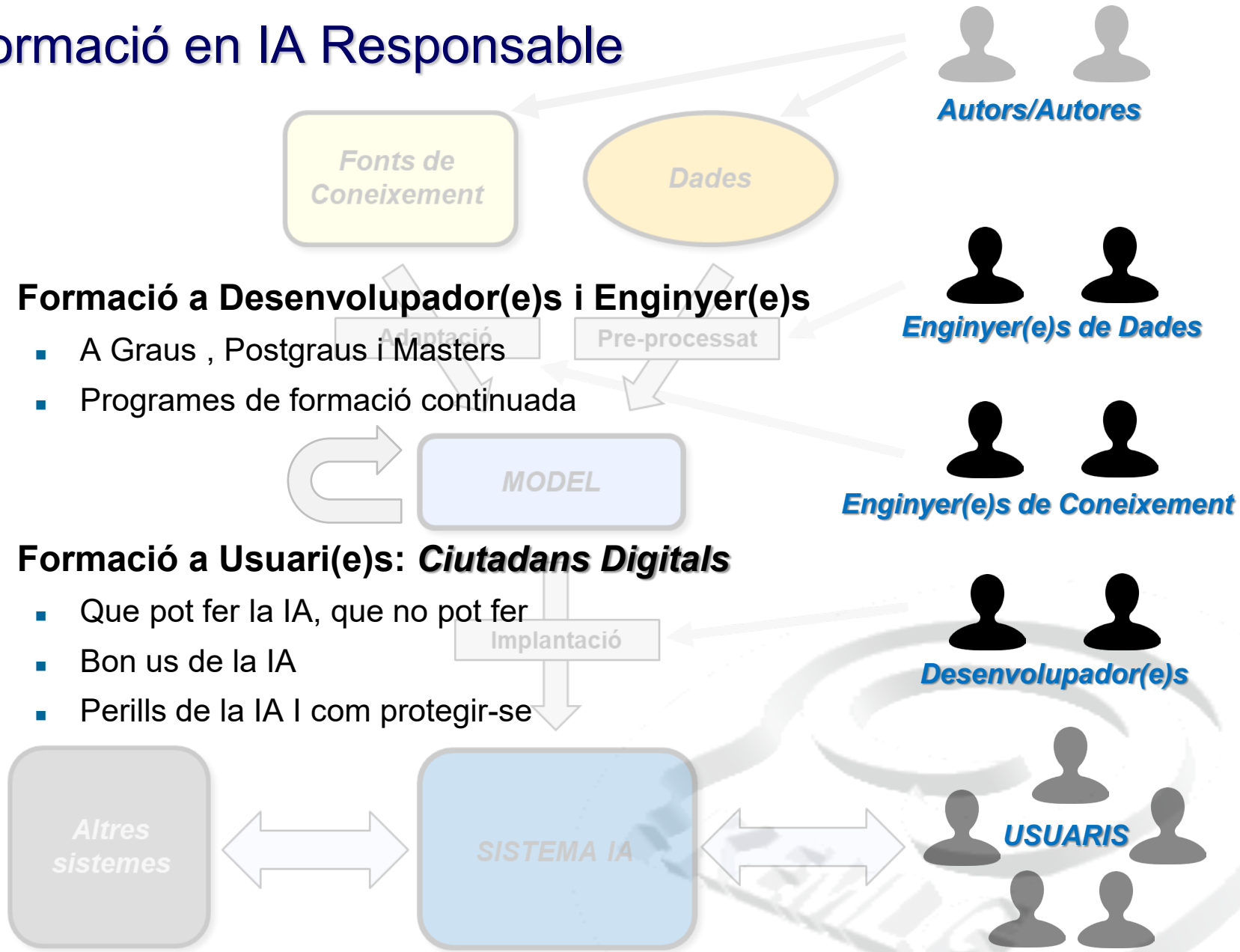
# Formació en IA Responsable

- **Formació a Desenvolupador(e)s i Enginyer(e)s**

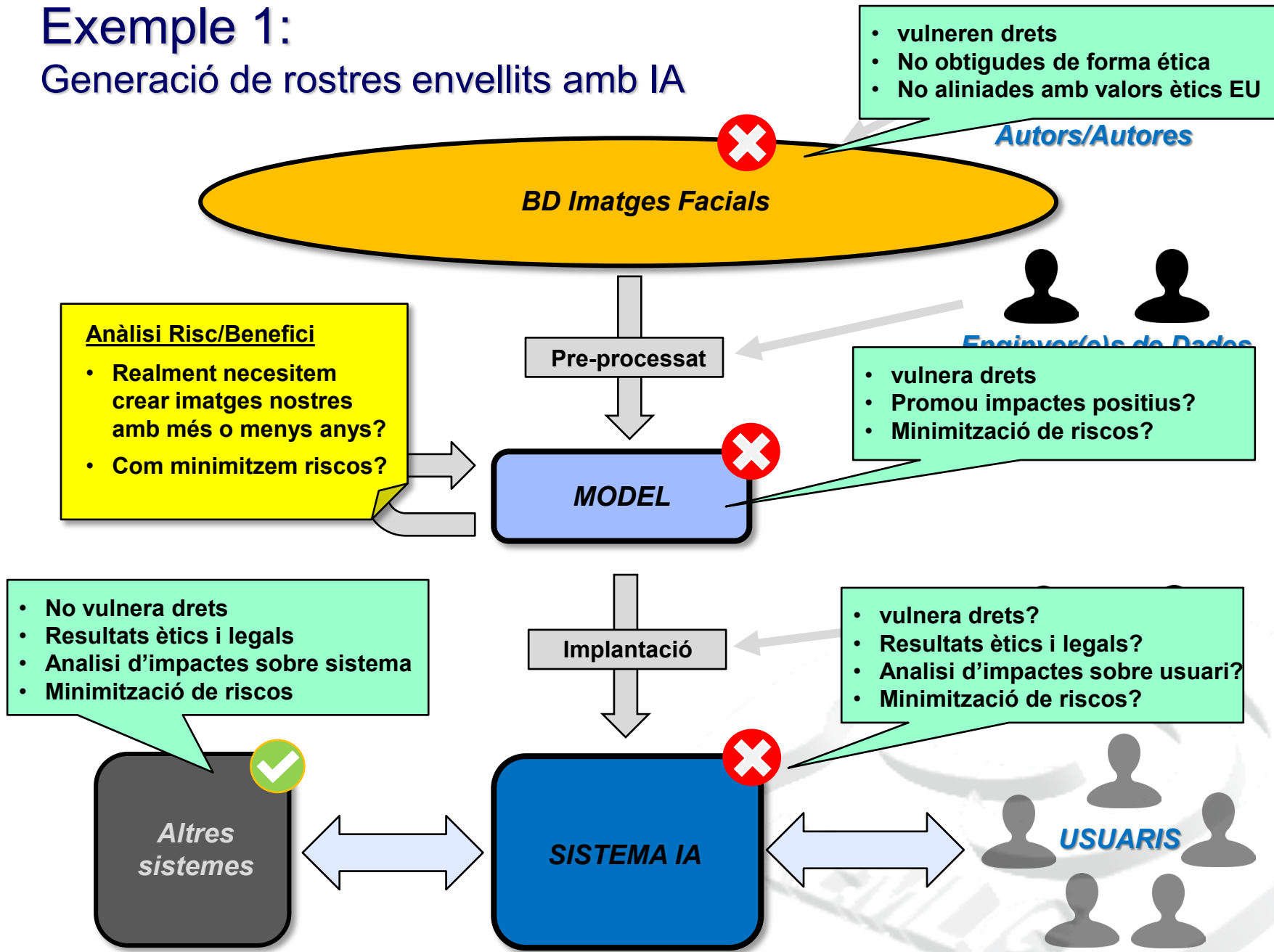
- A Graus , Postgraus i Masters
- Programes de formació continuada

- **Formació a Usuari(e)s: *Ciudadans Digitals***

- Que pot fer la IA, que no pot fer
- Bon us de la IA
- Perills de la IA i com protegir-se



# Exemple 1: Generació de rostres envellits amb IA



# Exemple 2: Ús de ChatGPT com servei

- No vulneren drets
- Obtingudes de forma ètica i legal
- Aliniades amb valors ètics

**Fonts de Coneixement** ✓

**Dades** ✓

- No vulneren drets
- Obtingudes de forma ètica i legal
- Aliniades amb valors ètics

*Autors/Autores*

**Anàlisi Risc/Benefici**

- Realment necessitem ChatGPT? (Hi han moltes tecnologies de diàleg que si son Responsables)
- Com minimitzem riscos?

Adaptació

Pre-processat

- No vulnera drets
- Promou impactes positius
- Minimització de riscos

*Enginyer(e)s de Dades*

**MODEL** ✓

*Enginyer(e)s de Coneixement*

Implantació

- vulnera drets?
- Resultats ètics i legals?
- Anàlisi d'impactes sobre usuari?
- Minimització de riscos?

- Pot vulnerar drets?
- Resultats ètics i legals?
- Anàlisi d'impactes sobre sistema
- Minimització de riscos?

GPT3  
GPT4  
GPT5

**ChatGPT** ✗

**SISTEMA IA** ?

**USUARIS**

# Exemple 3: Reentrenament de (Chat)GPT i altres LLMs

- Vulneren drets?
- Obtigudes de forma ètica i legal?
- No aliniades amb valors ètics EU

- No vulneren drets
- Obtigudes de forma ètica i legal
- Aliniades amb valors ètics

LLaMa (filtrat)  
GPT3/GPT4/GPT5  
PaLM  
Bloom

Dades

Autors/Autores

**Anàlisi Risc/Benefici**

- Analitzar si el LLM que reentrem es Responsable
- Anàlisi molt acurat per minimitzar riscos

Reentrenament

Pre-processat

- vulnera drets?
- Promou impactes positius?
- Minimització de riscos?

Enginyer(e)s de Dades

MODEL

Enginyer(e)s de Coneixement

- No vulnera drets
- Resultats ètics i legals
- Anàlisi d'impactes sobre sistema
- Minimització de riscos

Implantació

- vulnera drets?
- Resultats ètics i legals?
- Anàlisi d'impactes sobre usuari?
- Minimització de riscos?

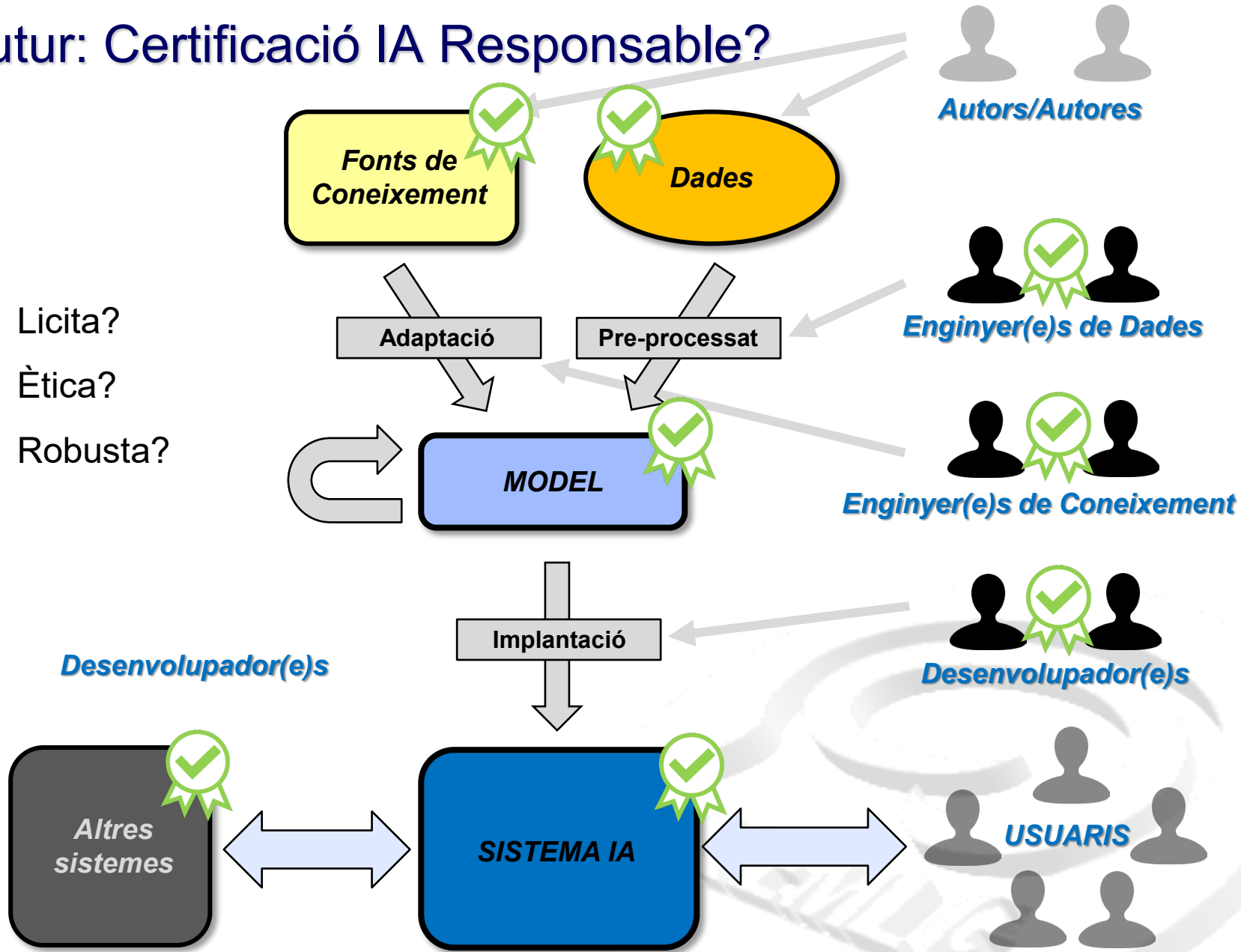
Altres sistemes

SISTEMA IA

USUARIS

# Futur: Certificació IA Responsable?

- Lícita?
- Ètica?
- Robusta?





# Conclusions

---



Knowledge Engineering and Machine Learning Group  
UNIVERSITAT POLITÈCNICA DE CATALUNYA

<https://kemlg.upc.edu>

informàTICa  
Enginyeria en informàtica  
de Catalunya

- **Hi ha IA més enllà de la IA Generativas**
  - Hi han bons i mals usos de la IA
  - Hi han IAs que son més fàcils de desenvolupar seguint els principis de la IA Responsable
- **La IA Generativa no es dolenta “per se”**
  - Però fins ara no s’ha desenvolupat amb els principis de IA Responsable
  - Possibles riscos per usuaris i desenvolupadors
- **No n’hi ha prou amb una Regulació de la IA**
  - Cal formar en els principis de la IA Responsable a desenvolupadors i usuaris
  - Si un usuari te coneixement per reconeixer el perill, el pot evitar

# *De la IA Generativa a la IA Responsable*

*Javier Vázquez-Salceda*

*Cap del grup de recerca KEMLG-IDEAI*

*Vicedegà de IA del Col·legi d'Enginyeria en Informàtica de Catalunya*

*Membre del Grup Intercol·legial de Treball en IA (GTAI)*

*[jvazquez@cs.upc.edu](mailto:jvazquez@cs.upc.edu)*

*<http://www.cs.upc.edu/~jvazquez>*



Knowledge Engineering and Machine Learning Group  
UNIVERSITAT POLITÈCNICA DE CATALUNYA

*<https://kemlg.upc.edu>*

informàTICa  
Enginyeria en informàtica  
de Catalunya