IA Generativa y LLMs para la Transformación Organizacional

FEN Unegocios

2025-11-24

Enlaces

- segana@fen.uchile.cl
- https://segana.netlify.app
- https://www.linkedin.com/in/sebastian-egana-santibanez/
- https://github.com/sebaegana

Clase 02

Objetivo general

- Analizar el rol de la inteligencia artificial generativa en la estrategia empresarial

---

Pregunta activadora

¿Qué procesos de su organización creen que podrían mejorar con IA generativa?

¿Qué es (y qué no es) Inteligencia Artificial?

- **IA:** capacidad de sistemas para aprender patrones y tomar decisiones basadas en datos
- **No es:** magia ni conciencia, sino algoritmos que **aprenden de ejemplos**

- Se basa en **estadística, aprendizaje automático y automatización**

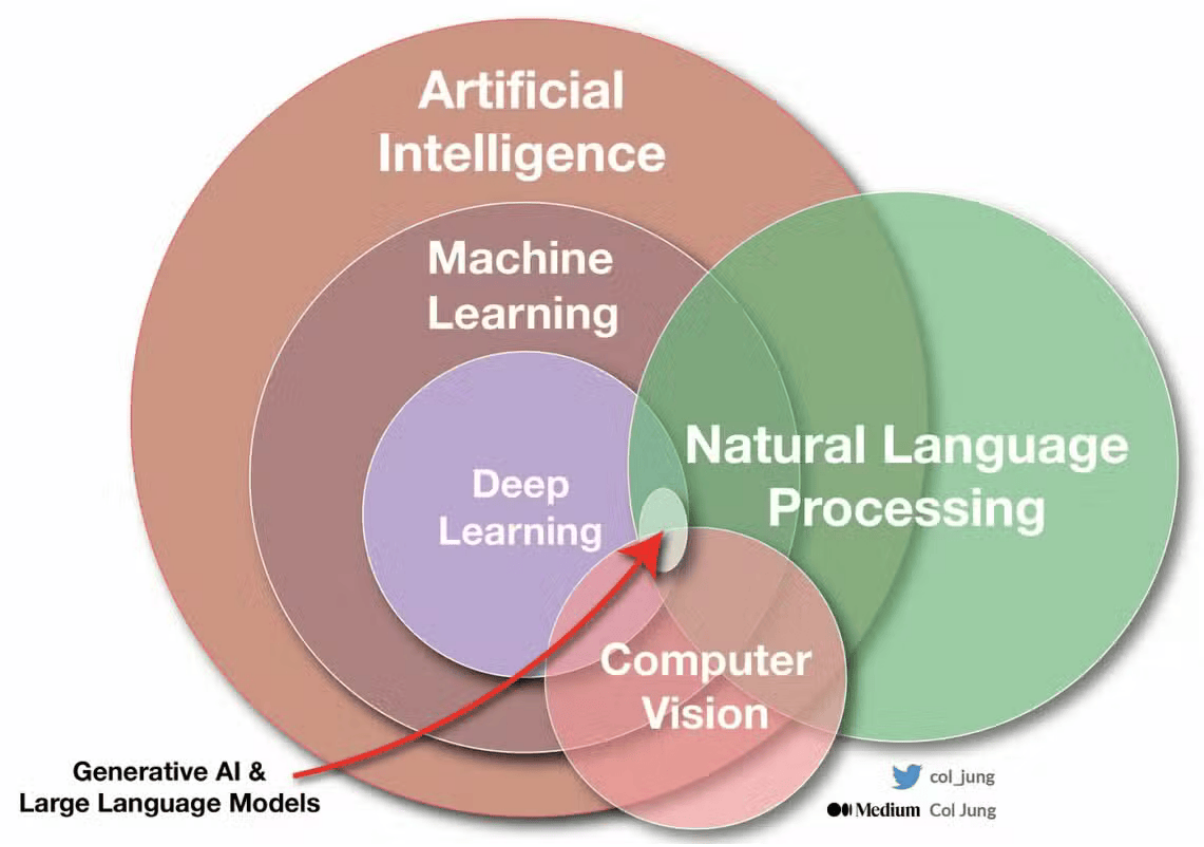


Figura 1: Recuperado en: https://www.orsys.fr/orsys-lemag/es/aprendizaje-automatico-aprendizaje-profundo-ia-diferencias

---

¿Qué es la IA Generativa?

Conceptos fundamentales

- **IA (Artificial Intelligence):** máquinas que realizan tareas que suelen requerir inteligencia humana.

- **IA Generativa:** sistemas capaces de *crear* contenido como lo hacemos los humanos (texto, imágenes, audio, código).
- Idea central: La IA generativa predice patrones para producir contenido original.
- IA y algoritmos: Algoritmos como serie de reglas dadas; en el caso de la IA esos algoritmos son construidos por la IA y puede ser mejorados.

Innovaciones clave en la historia

1. **Cloud computing** → datos + cómputo

2. **GANs** → imágenes fotorrealistas → Ver enlace
3. **Transformers** → comprensión profunda del lenguaje → Ver enlace
4. **RLHF** → modelos afinados con retroalimentación humana

5. **Midjourney / Stable Diffusion / GPT-4** → multimodalidad y creatividad avanzada

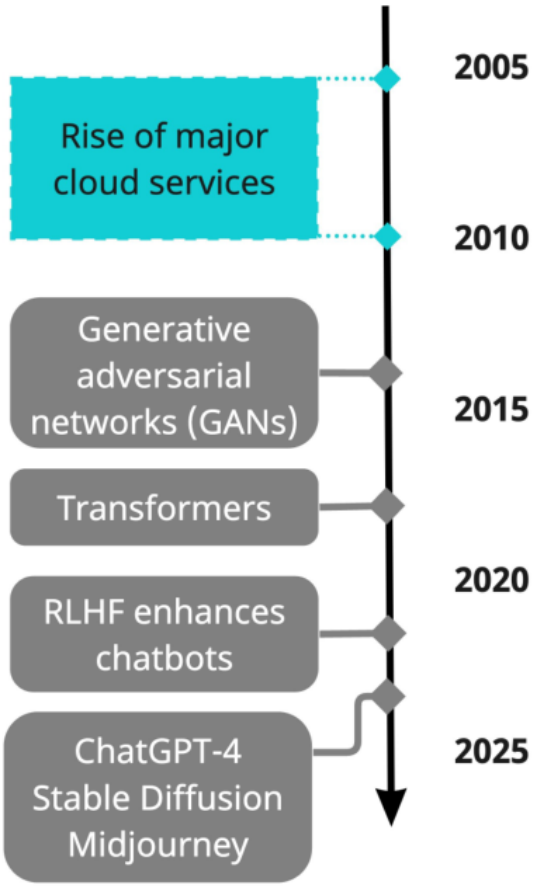


Figura 2: Recuperado de: Curso *Generative AI for Business*, DataCamp

---

Para revisar

> **i** Traditional AI vs. Generative AI: What's the Difference?
>
> To understand the benefits and uses of artificial intelligence (AI) in education, you first must discern the differences between traditional AI and generative AI. Here's a comparison of the two.
> Leer nota completa

La IA no es un modelo — es un ecosistema de componentes

Un **LLM** (Large Language Model) como GPT o Claude es solo una parte del sistema.

Para responder una consulta compleja, se articulan **varios segmentos de IA**:

- **Computer Vision** → detecta y analiza imágenes (objetos, rostros, escenas).
- **OCR (Reconocimiento Óptico de Caracteres)** → extrae texto de imágenes o PDFs.
- **Speech Recognition / TTS** → convierte audio en texto y viceversa.
- **LLM** → interpreta, contextualiza y genera una respuesta en lenguaje natural.
- **Reasoning / Orchestration Layer** → decide qué componentes usar y en qué orden.

- Resultado: "El documento corresponde a un contrato de servicios con fecha de inicio 2024-01-10 entre X y Y."

¿Qué es un LLM? (Large Language Model)

Un **LLM (Large Language Model)** es un modelo de inteligencia artificial entrenado con **enormes cantidades de texto** para:

- predecir la siguiente palabra
- generar respuestas coherentes
- comprender contextos complejos
- producir contenido (texto, código, instrucciones)
- razonar de manera estadística

Idea clave

Un LLM **no entiende**, sino que **modela patrones del lenguaje** para generar la respuesta más probable según los datos con los que fue entrenado.

¿Cómo procesan texto los LLMs?

Los modelos de lenguaje procesan texto mediante una serie de pasos y **parámetros clave** que controlan cómo entienden, representan y generan lenguaje.

Parámetros internos del modelo (no ajustables por el usuario)

Son parte de la arquitectura del LLM y definen cómo representa y procesa lenguaje.

- embedding dimension (p. ej. 768, 1536, 4096…)
- tokenizer (reglas que dividen el texto)
- context window (máximo de tokens simultáneos)
- número de capas (layers)
- número de cabezas de atención (attention heads)
- matrices de atención
- parámetros totales del modelo (ej. 8B, 70B, 1T)

Tokens

Los LLMs no leen palabras completas: dividen el texto en **tokens**, que pueden ser:

- sílabas
- fragmentos de palabra
- palabras completas (si son comunes)
- signos de puntuación

Ejemplos: "inteligencia" → "in", "tel", "igen", "cia"

Claves:

- Más tokens → más costo computacional
- Todos los modelos operan en *token space*, no en texto literal
- Todo input y output se mide en tokens

Embeddings

Cada token se transforma en un **vector numérico** en un espacio de cientos o miles de dimensiones.

Esto permite que el modelo codifique:

- significado
- contexto
- relaciones semánticas
- similitud entre conceptos

Los embeddings son el "idioma interno" del modelo.

Sobre los embeddings

Entrenamiento previo de los embeddings

Los embeddings que usamos (por ejemplo, con OpenAI o modelos open source) no se crean desde cero cada vez que los aplicamos. En realidad, provienen de un proceso de entrenamiento previo (pre-training) realizado sobre enormes cantidades de texto.

> **i** Getting Started With Embeddings
>
> We're on a journey to advance and democratize artificial intelligence through open source and open science.
> Leer nota completa

¿Qué aprende el modelo?

Durante el entrenamiento, el modelo analiza millones o miles de millones de frases y aprende relaciones estadísticas entre palabras. El objetivo es que, al ver una secuencia de texto, el modelo pueda predecir la siguiente palabra o reconocer palabras que encajan en el mismo contexto.

Texto de entrenamiento	Tarea implícita	Lo que aprende
"El **perro** ladra en el jardín."	Predecir "perro" a partir de su contexto ("El … ladra…")	Que "perro" aparece en contextos parecidos a "gato", "animal", "mascota".
"El **avión** aterriza en la pista."	Predecir "avión"	Que "avión" se asocia a "vuelo", "piloto", "aeropuerto".

Así, el modelo descubre automáticamente los significados y las similitudes semánticas sin reglas escritas por humanos.

¿Qué produce este aprendizaje?

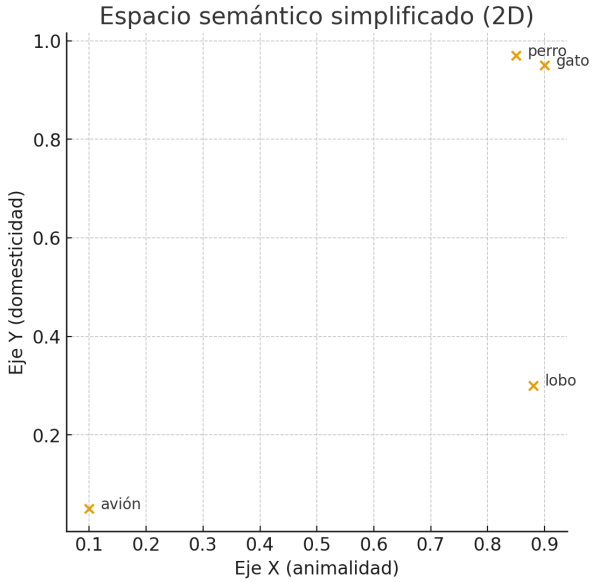
El modelo transforma cada palabra (o token) en un vector numérico que resume su significado aprendido. Estos vectores viven en un espacio de alta dimensión (por ejemplo, 768 o 1536 dimensiones). En ese espacio, las palabras que comparten contexto quedan cerca entre sí, y las que no tienen relación quedan lejos.

Ejemplo esquemático (espacio 2D simplificado):

Imaginemos que tenemos solo dos dimensiones (en realidad hay miles). Cada palabra se convierte en un punto:

Palabra	Eje X (animalidad)	Eje Y (domesticidad)
gato	0.9	0.95
perro	0.85	0.97
lobo	0.88	0.30
avión	0.10	0.05

En el gráfico:



- gato y perro están muy cerca, porque comparten significado.
- lobo está un poco más lejos: sigue siendo animal, pero no doméstico.
- avión está totalmente fuera del grupo

Un ejemplo en python:

Instalación de librerías:

```
pip install openai numpy faiss-cpu
```

Generación de embeds

```
from openai import OpenAI
import numpy as np
import faiss  # base de datos vectorial
client = OpenAI()

# Textos a convertir en embeddings
sentences = [
    "The dog is barking.",
    "A cat is sleeping on the couch.",
    "A car is driving down the street.",
    "I love my pet animal."
]

# Crear embeddings usando el modelo de OpenAI
response = client.embeddings.create(
    model="text-embedding-3-small",
    input=sentences
)

# Extraer los vectores
vectors = np.array([data.embedding for data in response.data])
print("Dimensión de los embeddings:", vectors.shape)
```

Salida esperada:

```
Dimensión de los embeddings: (4, 1536)
```

```
# Crear un índice FAISS
index = faiss.IndexFlatL2(vectors.shape[1])
index.add(vectors)

# Consultar algo nuevo
query = "My puppy is barking loudly."
query_vec = client.embeddings.create(
    model="text-embedding-3-small",
    input=query
).data[0].embedding

# Buscar el texto más parecido
D, I = index.search(np.array([query_vec]), k=2)
```

```
print("Resultados similares:")
for idx in I[0]:
    print("-", sentences[idx])
```

Resultados similares:

- The dog is barking.
- I love my pet animal.

Parámetros de entrenamiento (cómo se entrenan los modelos)

Estos parámetros son importantes para entender el "por qué" del desempeño de un LLM.

Tamaño del modelo (número de parámetros)

Los "parámetros" son los pesos internos del modelo (similares a neuronas).

- Llama 3: 8B–70B
- GPT-3: 175B
- GPT-4 real: desconocido (estimado entre 800B y 1.8T)
- Claude 3 Opus: no divulgado, estimado muy alto

Más parámetros mejor siempre

Depende del entrenamiento, datos y arquitectura.

Embedding dimension

Cantidad de dimensiones en que se representan los vectores.

- Modelos pequeños: 512–2048
- Modelos grandes: 4096–16384

Afecta:

- capacidad semántica
- riqueza de representación
- costo computacional

Batch size (en entrenamiento)

Cantidad de ejemplos procesados simultáneamente.

- batch grande → aprendizaje más estable
- batch pequeño → aprendizaje más ruidoso

Esto afecta:

- calidad final
- velocidad de entrenamiento
- requerimientos de GPU

(No confundir con "batch size" de inferencia en APIs.)

Longitud del pre-training

Los modelos mejoran no solo por tamaño, sino por:

- más pasos de entrenamiento
- mejores datos
- mejores instrucciones
- mejor RLHF

Ventana de contexto (context window)

Es la cantidad máxima de tokens que el modelo puede considerar **al mismo tiempo**.

- GPT-4: ~128k
- GPT-4o: ~200k
- Claude 3 Opus: ~200k
- Gemini 1.5 Ultra: ~1M
- Llama 3: 8k–128k (según versión)

Más contexto = modelos con "memoria a corto plazo" más amplia.

Parámetros de generación (cómo controlamos la salida)

Estos parámetros **no afectan cómo el modelo entiende**, sino **cómo genera texto**. Son esenciales en prompt engineering.

Temperatura (temperature)

Controla la **aleatoriedad**.

- Temperatura baja (0.0–0.3):
  - más precisión
  - respuestas estables y deterministas
- Temperatura alta (0.7–1.2):
  - más creatividad
  - más variación
  - mayor riesgo de errores

Regla:

- análisis → baja
- creatividad → alta

Top-p (nucleus sampling)

Controla el **porcentaje de probabilidad acumulada** desde donde el modelo puede elegir la próxima palabra.

- top-p = 0.1 → muy restrictivo
- top-p = 0.9 → muy diverso
- top-p = 1.0 → sin restricciones

Usado junto a temperatura para ajustar estilo y creatividad.

Top-k

Límite del número de tokens candidatos que el modelo puede elegir.

- top-k = 1 → modo determinista
- top-k = 40 → moderado
- top-k = 100+ → más creativo

Penalización de repetición (repetition penalty)

Evita que el modelo repita frases o palabras.

- Valores altos = menos repetición

Ideal para textos largos, código y redacción formal.

Max tokens (longitud máxima de respuesta)

Define cuántos tokens **puede generar como salida**.

Si max_tokens es muy pequeño → respuestas truncadas.
Si es grande → textos más detallados.

Resumen visual

Concepto	Rol
Token	Unidad mínima procesada
Embedding	Representación numérica del significado
Ventana de contexto	Memoria a corto plazo
Temperatura	Aleatoriedad
Top-p / top-k	Diversidad del texto
max_tokens	Longitud de salida
Parámetros (weights)	Capacidad del modelo
Embedding dimension	Profundidad conceptual
Batch size	Estabilidad durante entrenamiento

Frase clave:

> Los LLMs no "piensan": calculan probabilidades en un espacio matemático gigante lleno de significados.

Tipos de LLMs (con ejemplos)

Los modelos actuales pueden clasificarse en **cuatro grandes categorías**, según su entrenamiento y capacidades.

Modelos base (Base Models)

Entrenados con texto masivo, sin instrucciones humanas.
Generan lenguaje, pero no siguen órdenes de manera natural.

Ejemplos:

- **GPT-3** (OpenAI, 2020)
- **Llama 2 Base** (Meta, 2023)
- **Mistral Base 7B** (Mistral AI, 2023)
- **Gemma Base** (Google, 2024)
- **Falcon 40B Base** (Technology Innovation Institute)

Uso típico: pre-entrenamiento, fine-tuning, tareas no conversacionales.

Modelos Instruct (afinados para seguir instrucciones)

Entrenados con **RLHF + datasets de instrucciones**. Son los modelos "para conversar".

Ejemplos:

- **GPT-3.5 Turbo** (OpenAI)
- **GPT-4, GPT-4o** (OpenAI)
- **Claude 2, Claude 3 Sonnet / Opus** (Anthropic)
- **Gemini Advanced (1.5 Pro / Ultra)** (Google)

- **Llama 3 Instruct 8B / 70B** (Meta, 2024)
- **Mistral Instruct 7B / 8x22B Mixtral** (Mistral AI)
- **Command R / R+** (Cohere)
- **Qwen 2 Instruct** (Alibaba)

Uso típico: chat, preguntas/respuestas, análisis de texto, resúmenes, tareas empresariales.

Modelos multimodales (texto + imágenes + audio + video)

Pueden **ver, escuchar, interpretar y generar contenido** en múltiples formatos.

Ejemplos principales (2024–2025):

- **GPT-4o** (OpenAI) — visión + audio + razonamiento
- **Gemini 1.5 Pro / Ultra** (Google) — video + imágenes + razonamiento largo
- **Claude 3 Haiku / Sonnet / Opus** (Anthropic) — visión y análisis profundo
- **Llama 3 Vision** (Meta)
- **Mistral NeMo Vision** (Mistral AI)
- **Qwen2-VL** (Alibaba)
- **Pika Labs** (generación de video)
- **Runway Gen-2 / Gen-3** (video)

Uso típico: interpretación de imágenes, resúmenes de PDFs, análisis de video, agentes con audio.

Modelos especializados (fine-tuning o entrenamiento específico)

Optimizados para **una disciplina o tipo de tarea**.

Modelos para código

- **GPT-4o y GPT-4 Turbo (Code Interpreter / o1)**
- **DeepSeek-Coder**
- **StarCoder / StarCoder2**
- **Code Llama**
- **Phi-3 Mini para coding (Microsoft)**

Modelos para biología / ciencias

- **Evo (OpenAI)** – proteínas
- **AlphaFold 2 / AlphaFold 3 (DeepMind)**
- **ESM-2 (Meta)** – secuencias proteicas

Modelos para agentes / razonamiento profundo

- **OpenAI o1 / o3**
- **Gemini 1.5 Ultra (Chain-of-thought avanzado)**
- **ReAct + modelos instruct** (framework)

Modelos para finanzas / negocios

- **BloombergGPT**
- **FinGPT**
- **Claude 3 Opus** (muy fuerte en análisis estructurado)
- **Llama 3 + fine-tuning sectorial** (muy común en empresas)

Modelos para idiomas específicos

- **Yi (China)**
- **Qwen 2 (China)**
- **Mistral Small (UE)**
- **LLaMA 3 Spanish fine-tuned**

Modelos pequeños y rápidos (edge)

- **Phi-3 (Microsoft)** — <4B parámetros
- **LLaMA 3.1 8B**
- **Mistral 7B**
- **Gemma 2B / 7B**

Uso típico: casos de uso específicos, empresas que necesitan privacidad, modelos on-premise o en edge devices.

BONUS: Modelos de imágenes, audio y video (no LLM pero complementarios)

- **DALL · E 3** (OpenAI)
- **Midjourney v6**
- **Stable Diffusion XL / 3.0**
- **Suno v3 (audio/música)**
- **ElevenLabs (voz)**
- **Runway Gen-2 / Gen-3 (video)**

Permiten construir sistemas multimodales completos.

Cómo funciona: entrenamiento y prompting

- Entrenamiento: El modelo aprende patrones desde grandes volúmenes de datos.
- Prompting: El usuario da una instrucción → el modelo genera contenido.
- Ejemplo: Prompt: *"Escribe un haiku sobre IA generativa."*

  Un prompt no es solo una pregunta: es un programa breve que guía el comportamiento del modelo.

Interpretación de resultados de IA

Qué significa un "resultado confiable"

- Un modelo **no entrega certezas**, entrega **probabilidades**
- La confiabilidad depende de calidad de datos, validación del modelo, comparación entre predicciones y realidad
- Ejemplo: *modelo de churn predice 80 % → no dice "renunciará", sino "probabilidad alta de egreso"*

Modelos discriminativos vs generativos

- **Discriminativos:** clasifican o predicen.
- **Generativos:** crean contenido.

Analogía:

- Discriminativo → "¿Es fraude o no?"
- Generativo → "Resume este caso de fraude."

Aplicaciones de negocio

- Atención al cliente
- Desarrollo de software
- Diseño de productos
- Documentación y reportes
- Automatización de tareas repetitivas
- Insights desde grandes volúmenes de texto

Rol estratégico de la IA Generativa

¿Reemplazo o colaboración?

Caso: Kasparov vs Deep Blue

> **Kasparov vs. Deep Blue | The Match That Changed History**
>
> Over 20 years ago, World Champion Garry Kasparov took on IBM and the super-computer Deep Blue in the ultimate battle of man versus machine. This was a monumental moment in chess history and was followed closely around the world. This match appealed to chess players, scientists, computer experts, and...
> Leer nota completa

Caso: AlphaGo vs Lee

> **AlphaGo vs. Lee: la máquina venció al humano - BBC News Mundo**
>
> El programa informático AlphaGo, diseñado por DeepMind para jugar el juego del Go, derrotó al mejor jugador del mundo, el surcoreano Lee Se-dol. Los expertos aseguran que es un hito para la inteligencia artificial.
> Leer nota completa

Caso: OpenAI Five vs Team OG

Explained Simply: How A.I. Defeated World Champions in the Game of Dota 2

In 2019, the world of esports changed forever. For the first time, a superhuman AI program learned to cooperate with copies of itself and…
Leer nota completa

Conclusión:
> La IA reemplaza tareas, no profesiones completas.

Transformación Organizacional con IA Generativa

La IA generativa no es solo una herramienta tecnológica: es un catalizador de transformación organizacional que impacta procesos, personas, cultura, modelos de negocio y estrategias de datos.

¿Por qué la IA transforma organizaciones?

La IA generativa impulsa cambios profundos porque:

- Aumenta capacidades humanas.
- Automatiza tareas de alto costo cognitivo.
- Reduce tiempos entre idea → prototipo → ejecución.
- Abre nuevas posibilidades de productos y servicios.
- Permite rediseñar procesos completos.

Idea clave: La pregunta ya no es *"¿podemos usar IA?"*, sino **"¿cómo reorganizamos nuestro trabajo para aprovechar IA de forma estratégica?"**

Alineamiento estratégico

La IA solo genera valor si está vinculada a metas del negocio.

Principios:

- Conectar iniciativas de IA con KPIs reales (eficiencia, ingresos, retención, calidad).
- Priorizar casos de uso de alto impacto y bajo riesgo.
- Evitar pilotos aislados o experimentación sin dirección.
- Diseñar una hoja de ruta que avance junto a la estrategia general de la organización.

Mensaje clave:

> La IA es una herramienta estratégica, no un juguete tecnológico.

Cambio cultural y alfabetización en IA

El mayor freno a la adopción no es técnico: es cultural.

Cambios necesarios:

- Pasar de "trabajo manual" a "trabajo aumentado por IA".
- Fomentar una mentalidad de experimentación.
- Entrenar a los equipos en pensamiento crítico aplicado a IA.
- Normalizar el uso cotidiano de modelos generativos como un nuevo integrante del equipo.

Alfabetización en IA:

- Uso efectivo de prompts.
- Conocer límites y riesgos del modelo.
- Evaluar la calidad de las respuestas.
- Integrar IA en tareas diarias sin perder criterio profesional.

Gobernanza y gestión del riesgo

La IA necesita un marco formal de gestión responsable:

Dimensiones de gobernanza:

- **Seguridad y privacidad de datos.**
- **Mitigación de sesgos y fairness.**
- **Transparencia en procesos y decisiones.**
- **Regulación y compliance.**
- **Políticas de uso responsable.**
- **Evaluación del impacto ecológico asociado al cómputo.**

Riesgos a gestionar:

- Alucinaciones.
- Sesgos inesperados.
- Ciberseguridad.
- Mal uso (deepfakes, automatizaciones sin control).
- Riesgos reputacionales.

Impacto transversal en el negocio

La IA generativa afecta múltiples áreas al mismo tiempo:

Marketing

- Personalización.
- Generación de contenido.
- Segmentación inteligente.

Ventas

- Redacción asistida.
- Score de oportunidades.
- Eficiencia comercial.

Servicio al cliente

- Chatbots aumentados.
- Resúmenes de interacciones.
- Derivación inteligente.

Operaciones / Supply Chain

- Predicción.
- Optimización.
- Automatización de tareas repetitivas.

Recursos Humanos

- Redacción de perfiles.
- Capacitación.
- Evaluación y retroalimentación más eficiente.

Identificación de oportunidades de IA

Las oportunidades pueden clasificarse en tres arquetipos:

1) Augmentation

La IA ejecuta parte de la tarea y el humano sigue controlando.

2) Co-creation

Humano + IA colaboran para producir un resultado final.

3) Replacement

La IA automatiza una tarea o proceso completo.

Criterios para detectar oportunidades:

- ¿La tarea es repetitiva?
- ¿Consume tiempo excesivo?
- ¿Implica generar contenido o analizar información?
- ¿La calidad del output se puede verificar fácilmente?
- ¿Hay potencial de mejorar costo, calidad o velocidad?

Framework operativo para transformar con IA

1) Mapear procesos

- Identificar tareas de alto costo cognitivo.
- Documentar pasos actuales.
- Detectar "IA-fit" en microtareas específicas.

---

2) Experimentar ("Trial & Iterate")

- Probar IA en tareas acotadas.
- Medir impacto real.
- Ajustar prompts, flujos y procesos.

3) Escalar

- Entrenar más equipos.
- Integrar IA en sistemas existentes.
- Mantener mejora continua con métricas claras.

La IA redistribuye capacidades dentro de la organización

La IA cambia **qué tareas hacen las personas** y **cómo se distribuye el trabajo**:

- Más tiempo para creatividad, análisis y decisiones.
- Menos tiempo en tareas repetitivas.
- Incremento de autonomía en roles operativos.
- Reorganización de funciones tradicionales (marketing, finanzas, CX, TI).
- Emergencia de nuevos roles:
  - AI Product Owner
  - Prompt Designer
  - Model Steward
  - AI Trainer

Ética, privacidad y riesgos en IA Generativa

La IA generativa introduce enormes oportunidades, pero también dilemas éticos que deben gestionarse desde una mirada estratégica, responsable y centrada en las personas.

Sesgos (Bias)

Los sesgos son uno de los riesgos más frecuentes y peligrosos en sistemas de IA. La IA aprende de datos históricos que pueden contener desigualdades y, por tanto, puede amplificarlas.

Por qué ocurre

- Datos desbalanceados o incompletos.
- Inclusión directa o indirecta de variables sensibles.
- Modelos entrenados sin supervisión ética.
- Falta de representatividad en los datos.

Cómo detectarlo

- Validar desempeño en múltiples subgrupos.
- Tests de fairness (como demographic parity).
- Auditorías independientes.
- Monitoreo continuo.

Estrategias de mitigación

- Remover variables sensibles.
- Rebalanceo o enriquecimiento de datasets.
- Supervisión humana.
- Incluir fairness desde el diseño, no solo en el despliegue.

Privacidad y el "privacy–personalization paradox"

La personalización aumenta la utilidad del sistema, pero genera tensiones con la privacidad. A mayor personalización, mayor necesidad de datos personales.

Riesgos

- Exposición de datos sensibles.
- Uso de información sin consentimiento.
- Datos enviados a terceros sin control.
- Incorporación no autorizada de datos en el entrenamiento.

Mitigaciones

- Políticas de transparencia y consentimiento claras.
- Minimización de datos.
- Encriptación, anonimización y retención limitada.
- Cumplimiento normativo (GDPR, LGPD, Ley 19.628).

Copyright y propiedad intelectual

La IA generativa plantea desafíos legales y éticos en torno a la autoría y el uso de datos.

Cloudflare CEO warns AI is keeping businesses and their customers apart - The Logic

Matthew Prince is calling on publishers and companies to block AI developers from taking their content, in hopes of forcing them to pay for it
Leer nota completa

Temas clave

- Autoría del contenido generado.
- Uso de materiales protegidos en entrenamiento.
- Derechos de imagen.
- Posible infracción involuntaria por parte del usuario.

Mitigaciones

- Claridad en los términos de uso del modelo.
- Modelos privados o on-premise para datos sensibles.
- Gestión responsable de datasets.
- Documentación sobre fuentes y licencias.

Transparencia, cajas negras y explicabilidad (XAI)

Muchos modelos de IA —especialmente los LLMs— funcionan como **cajas negras**: sabemos qué entra y qué sale, pero no cómo se tomó la decisión.

Riesgos

- Falta de confianza del usuario final.
- Dificultad para auditar o corregir errores.
- Riesgo regulatorio.
- Falta de trazabilidad en decisiones críticas.

Soluciones

- **Explainable AI (XAI):**
  - **LIME (Local Interpretable Model-agnostic Explanations):** explicaciones locales simplificadas.
  - **SHAP (SHapley Additive exPlanations):** importancia de variables para determinar la predicción.
- Uso de modelos interpretables cuando sea posible.
- Documentación de supuestos, límites y fuentes de datos.
- Transparencia en el ciclo de vida completo del modelo.

Accountability (Responsabilidad humana)

La IA no es una excusa para evitar responsabilidad. Siempre debe haber un responsable humano por las decisiones donde la IA participa.

> ℹ "Quería que ChatGPT me ayudara. Entonces, ¿por qué me aconsejó cómo suicidarme?" - BBC News Mundo
>
> ChatGPT escribió una nota de suicidio para una joven con problemas mentales y le aconsejó cómo quitarse la vida, según una investigación de la BBC.
> Leer nota completa

> OpenAI estima que más de un millón de sus 800 millones de usuarios semanales parecen expresar pensamientos suicidas

---

Problemas típicos

- Nadie asume responsabilidad por errores del modelo.
- "Lo dijo la IA" como racionalización.
- Automatización excesiva sin supervisión.
- Riesgo de confianza ciega en sistemas imperfectos.

Mitigaciones

- Roles claros (equipo de datos, producto, legal).
- Supervisión humana obligatoria ("human-in-the-loop").
- Procedimientos de revisión y auditoría.
- No delegar decisiones críticas exclusivamente a la IA.

Fairness y justicia algorítmica

El principio de **fairness** busca evitar discriminaciones directas o indirectas.

Dimensiones típicas

- Raza y etnia.
- Género.
- Edad.
- Nivel socioeconómico.
- Lenguaje, acento o país de origen.
- Discapacidad.

Ejemplos

- Modelos clínicos que priorizan ciertos subgrupos.
- Motores crediticios menos accesibles para minorías.
- Clasificadores de CVs entrenados con historiales sesgados.

Riesgos emergentes

Deepfakes y manipulación

- Contenido sintético difícil de distinguir de lo real.
- Riesgo reputacional.
- Manipulación política o social.
- Fraudes mediante audio y video falsificado.

Misinformación

- Generación masiva de noticias falsas.
- Ataques de saturación informativa.
- Erosión de la confianza pública.

Autonomía y control humano

- Sistemas que actúan fuera de los límites esperados.
- Dilema autonomía–control.
- Proyectos con potencial riesgo físico (autos autónomos, robots).

Ethical AI by Design (Ética por diseño)

Los principios éticos deben integrarse desde la concepción del proyecto, no como un "parche" posterior.

Principios clave

- Definir objetivos explícitos.
- Alinear con stakeholders relevantes.
- Gestionar datos con seguridad y criterios éticos.
- Diseñar con transparencia.
- Evaluar y mitigar sesgos desde el inicio.
- Abordar preocupaciones de impacto social.
- Iterar y monitorear continuamente.

Beneficios

- Menor riesgo legal y reputacional.
- Mayor adopción y confianza.
- Procesos de IA más robustos y auditables.
- Fomenta la innovación responsable.

¿Por qué importa todo esto?

- Aumenta la confianza del público y los usuarios.
- Reduce riesgos legales, operacionales y reputacionales.
- Fomenta la innovación sostenible.
- Permite aprovechar el potencial de la IA sin comprometer valores éticos.
- Convierte la ética en una ventaja competitiva.

La ética en IA no es solo buena práctica: es buena estrategia de negocios.

Referencias

- **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.** (2017). *Attention Is All You Need.* Advances in Neural Information Processing Systems (NeurIPS).
  Disponible en: https://arxiv.org/abs/1706.03762
- **OpenAI et al.** (2019). *Dota 2 with Large-Scale Deep Reinforcement Learning.* arXiv preprint arXiv:1912.06680.
  Disponible en: https://arxiv.org/pdf/1912.06680