

INTELLIGENZA ARTIFICIALE GENERATIVA

COS'È? COME FUNZIONA? PROBLEMI?!

Fonti:

- *Towards Data Science*
- *Grace Browne (Wired)*
- *IBM: Ai Bias*
- *Wikipedia: AI slop*
- *Articoli presenti nelle didascalie delle immagini*

MODELLI TRASFORMATIVI

DEFINIZIONE

Una rete neurale artificiale (**NN**) è un modello computazionale ispirato alla struttura e al funzionamento del cervello umano. È composta da unità chiamate **neuroni artificiali** organizzati in strati, che possono prendere in input dei numeri e possono emettere in output altri numeri.

Il **Deep learning** è un sottoinsieme del **Machine Learning (ML)** basato su livelli di reti neurali e sta alla base dell'intelligenza artificiale generativa.

Curiosità

il **Mark I Perceptron** del 1958

ESEMPIO: CLASSIFICARE UN OGGETTO

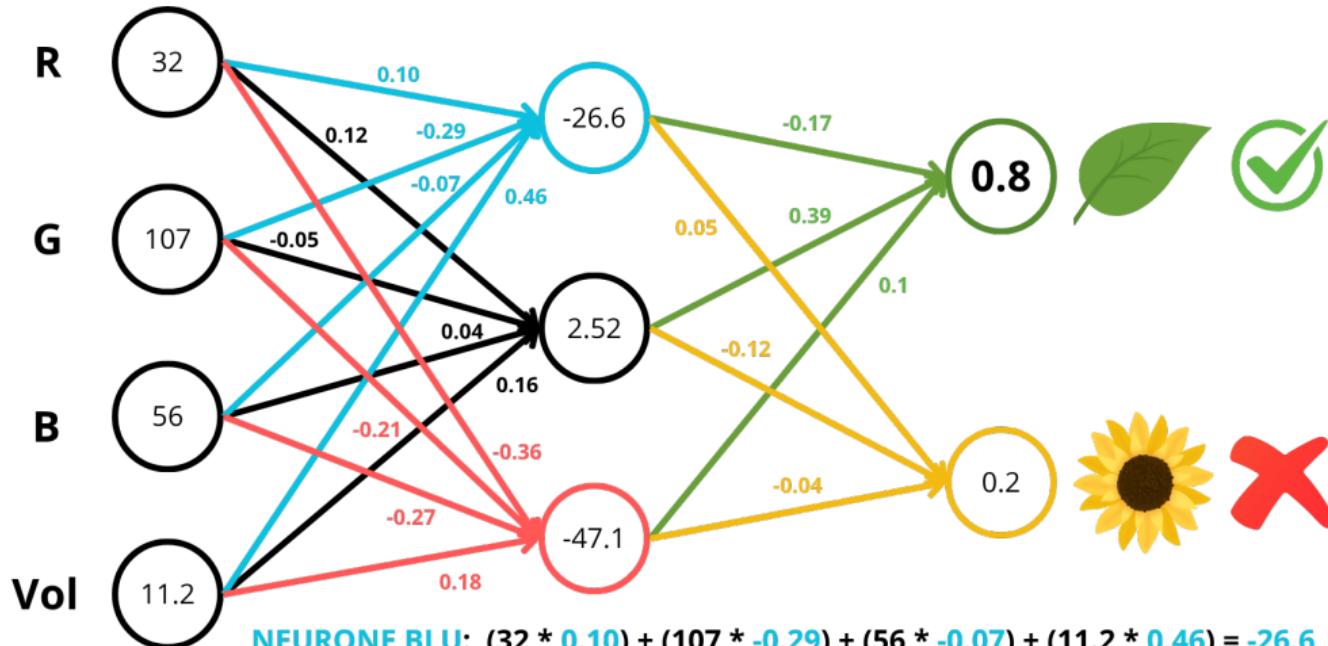
DATI OGGETTO DISPONIBILI:

- Colore dominante (RGB);
- Volume in millilitri.

Esempio: Dati di una foglia e di un girasole:

	Foglia	Fiore
R	32	241
G	107	200
B	56	4
Vol	11.2	59.5

ESEMPIO: CLASSIFICARE UN OGGETTO (FOGLIA O FIORE?)



$$\text{NEURONE BLU: } (32 * 0.10) + (107 * -0.29) + (56 * -0.07) + (11.2 * 0.46) = -26.6$$

$$\text{NEURONE ROSSO: } (32 * -0.36) + (107 * -0.21) + (56 * -0.27) + (11.2 * 0.18) = -26.6$$

$$\text{FOGLIA: } (-26.6 * -0.17) + (2.52 * 0.39) + (-47.1 * 0.1) = 0.8$$

$$\text{FIORE: } (-26.6 * 0.05) + (2.52 * -0.12) + (-47.1 * -0.04) = 0.2$$

Figura 1: creata con Canva

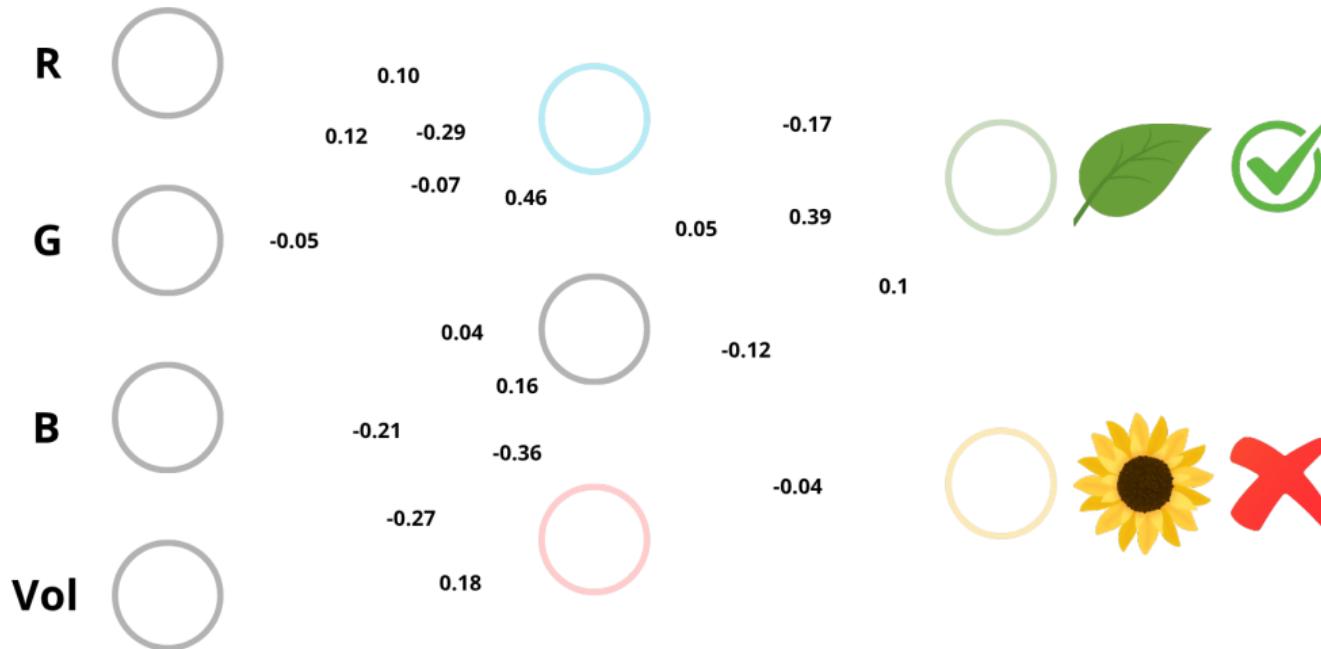
DEFINIZIONI

ELEMENTI DI UNA RETE NEURALE



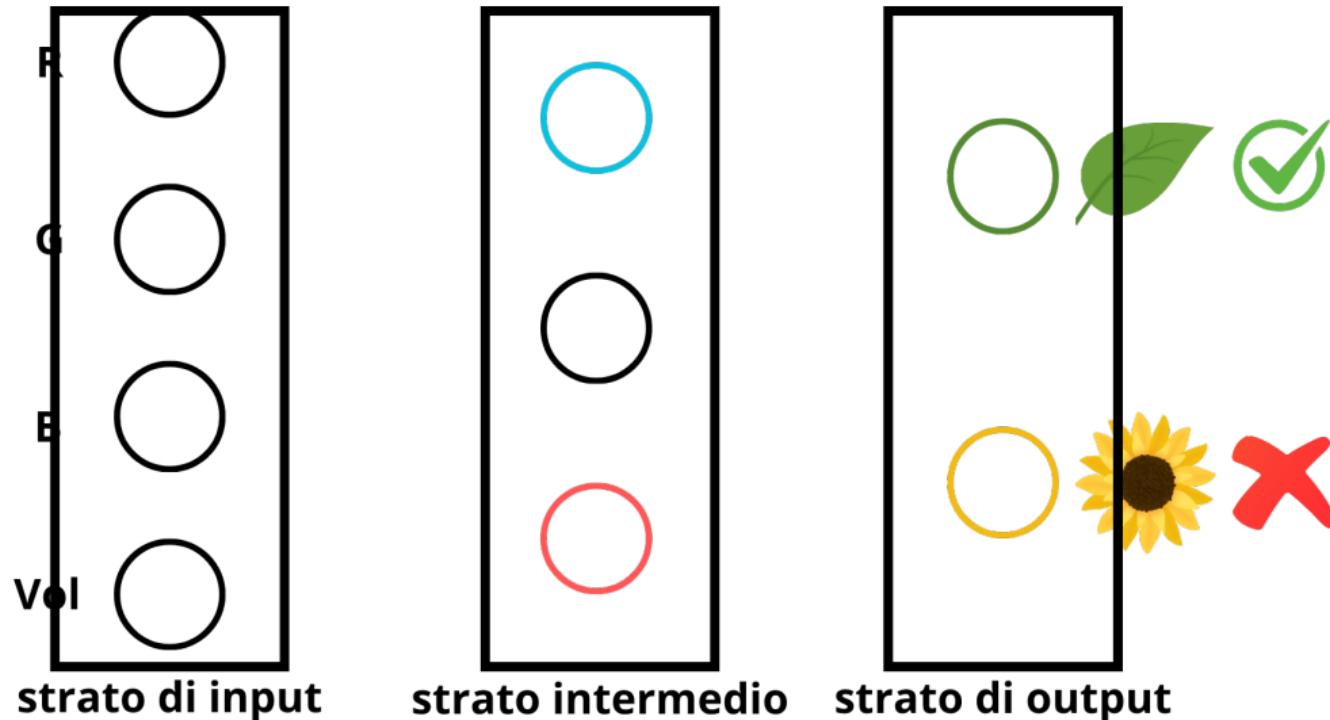
NEURONI/NODI

ELEMENTI DI UNA RETE NEURALE



PESI

ELEMENTI DI UNA RETE NEURALE



STRATI

ESEMPIO: CLASSIFICARE UN OGGETTO (FOGLIA O FIORE?)

INTELLIGENZA?!

Il modello **non ha idea** di cosa sia una foglia o un fiore, o cosa siano (RGB, Vol).
Il modello ha semplicemente il compito di prendere esattamente 4 numeri e restituire esattamente 2 numeri.

L'INTERPRETAZIONE DEI DATI IN INPUT E OUTPUT SPETTA A NOI.

COME SI DETERMINANO I PESI? ADDESTRARE IL MODELLO

COME SI DETERMINANO I PESI?

TRAINING DATA

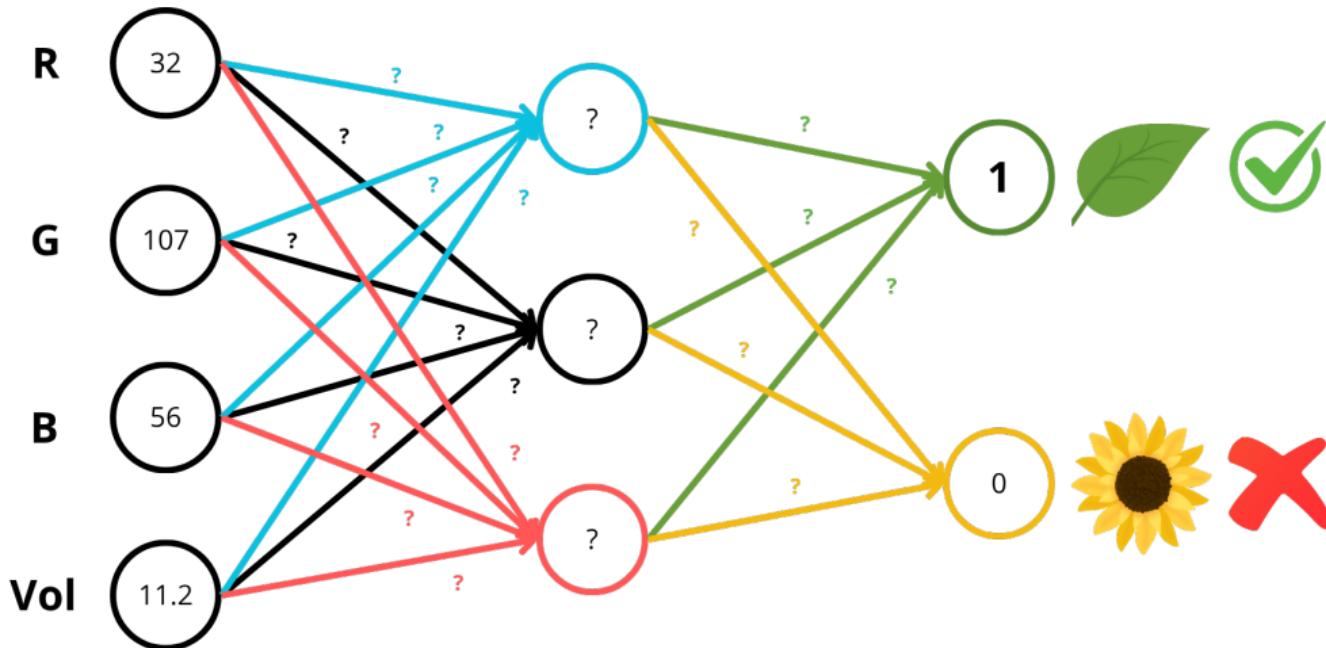
Per poter addestrare il modello e quindi determinare i pesi associati alle connessioni tra i neuroni, è necessario fornire al modello un insieme di dati di addestramento (**training data**), i quali contengono esempi di input e i corrispondenti output desiderati.

I dati vengono quindi **etichettati** in modo che il modello possa apprendere la relazione tra input e output.

Nell'esempio precedente i dati di addestramento etichettati potrebbero essere:

- Input: [32, 107, 56, 11.2] → Output desiderato: foglia;
- Input: [241, 200, 4, 59.5] → Output desiderato: fiore.

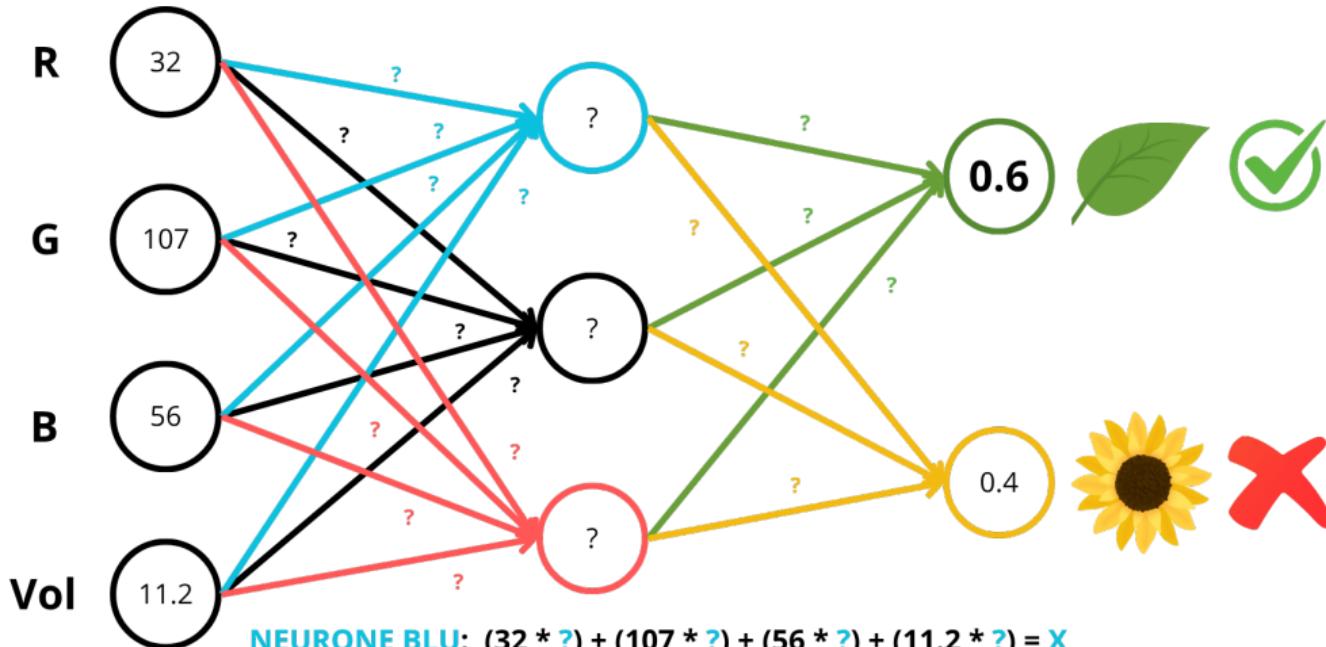
ALGORITMO DI APPRENDIMENTO



? = INIZIALMENTE VALORI CASUALI

Figura 3: creata con Canva

ALGORITMO DI APPRENDIMENTO



NEURONE BLU: $(32 * ?) + (107 * ?) + (56 * ?) + (11.2 * ?) = X$

NEURONE ROSSO: $(32 * ?) + (107 * ?) + (56 * ?) + (11.2 * ?) = Y$

FOGLIA: $(X * ?) + (? * ?) + (Y * ?) = 0.6$

FIORE: $(X * ?) + (? * ?) + (Y * ?) = 0.4$

Figura 3: creata con Canva

ALGORITMO DI APPRENDIMENTO

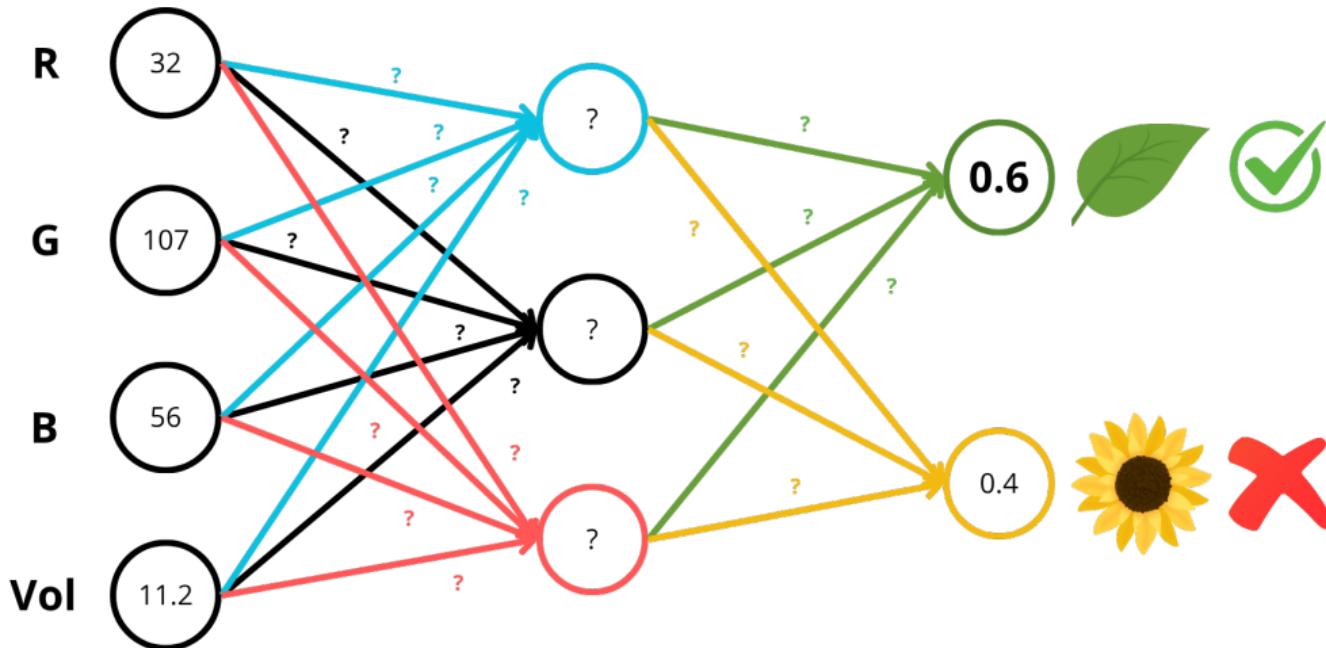
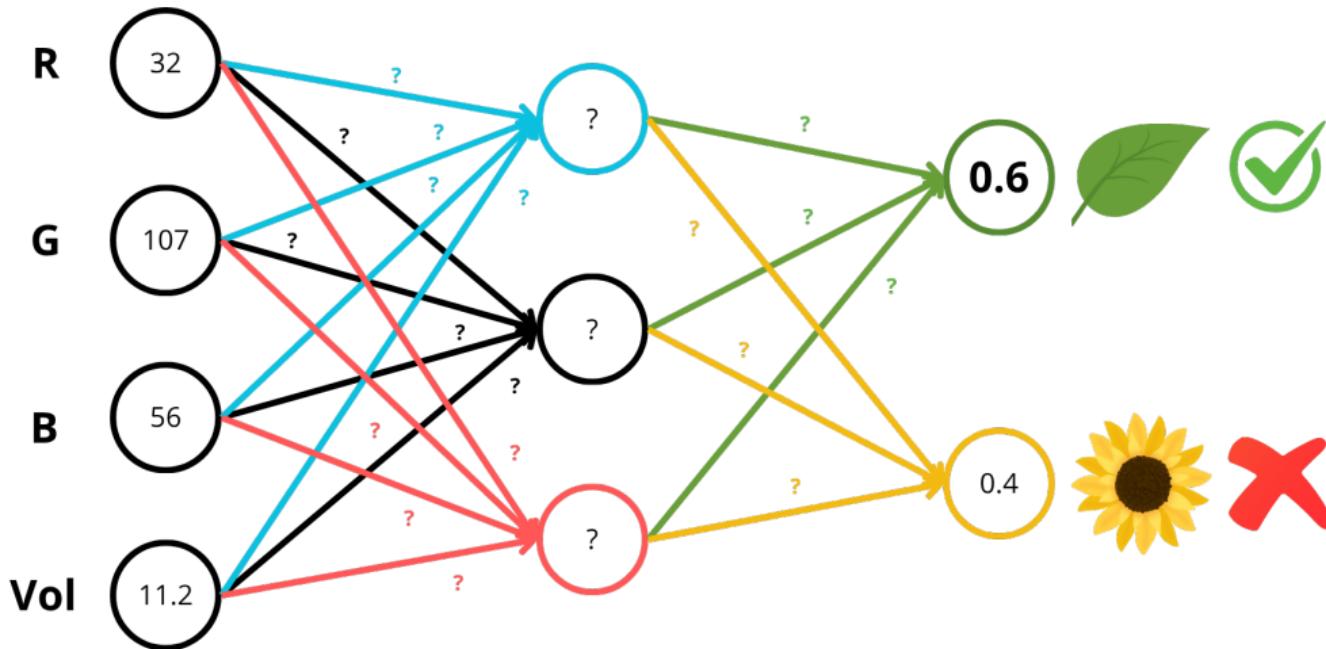


Figura 3: creata con Canva

ALGORITMO DI APPRENDIMENTO



$$\text{PERDITA} = (|1 - 0,6|) + (|0 - 0,4|) = 0,4 + 0,4 = \mathbf{0,8}$$

MINIMIZZARE LA PERDITA!

Figura 3: creata con Canva

ALLENAMENTO DEL MODELLO

Data la perdita, l'algoritmo di apprendimento utilizza un metodo chiamato **discesa del gradiente** per aggiornare i pesi della rete neurale. L'obiettivo è minimizzare la perdita, ovvero ridurre la differenza tra l'output previsto e l'output desiderato.

Questo processo viene ripetuto per **molte iterazioni**, utilizzando diversi esempi di dati di addestramento (**epoch**), fino a quando la rete neurale non raggiunge un livello accettabile di accuratezza.

In pratica, allenare le reti profonde è un processo duro e complesso perché i gradienti possono facilmente sfuggire al controllo, andando a zero o all'infinito durante l'allenamento. Con modelli moderni contenenti miliardi di parametri e funzioni molto più complesse, l'addestramento di un modello richiede enormi risorse di calcolo.

Curiosità

[Understanding LLMs from Scratch Using Middle School Math](#)

BLACK BOX

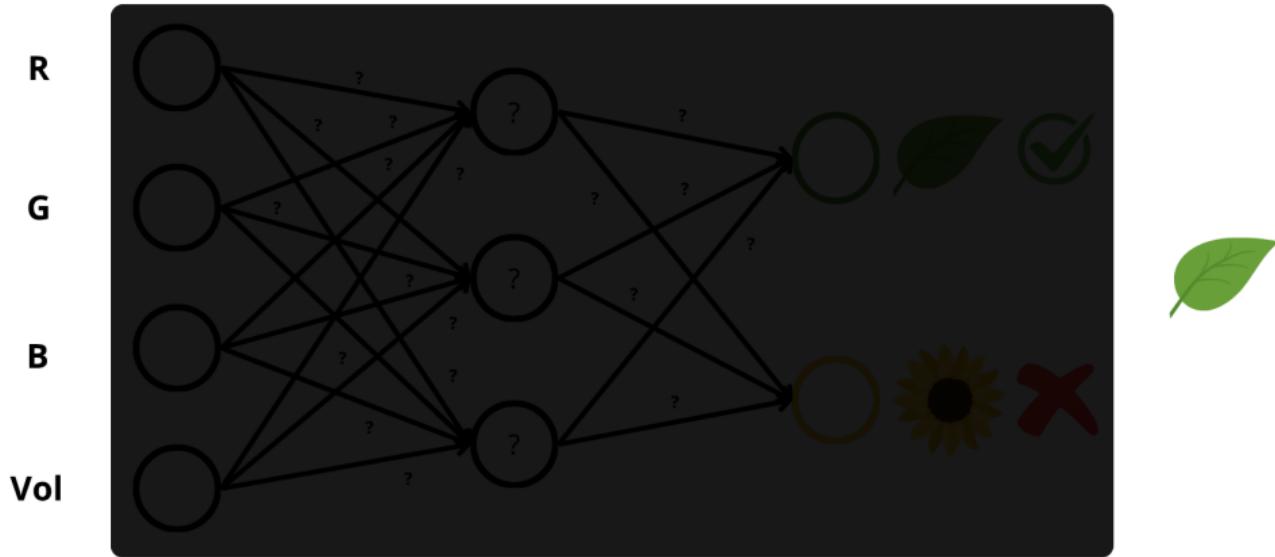


Figura 4: creata con Canva

ChatGPT

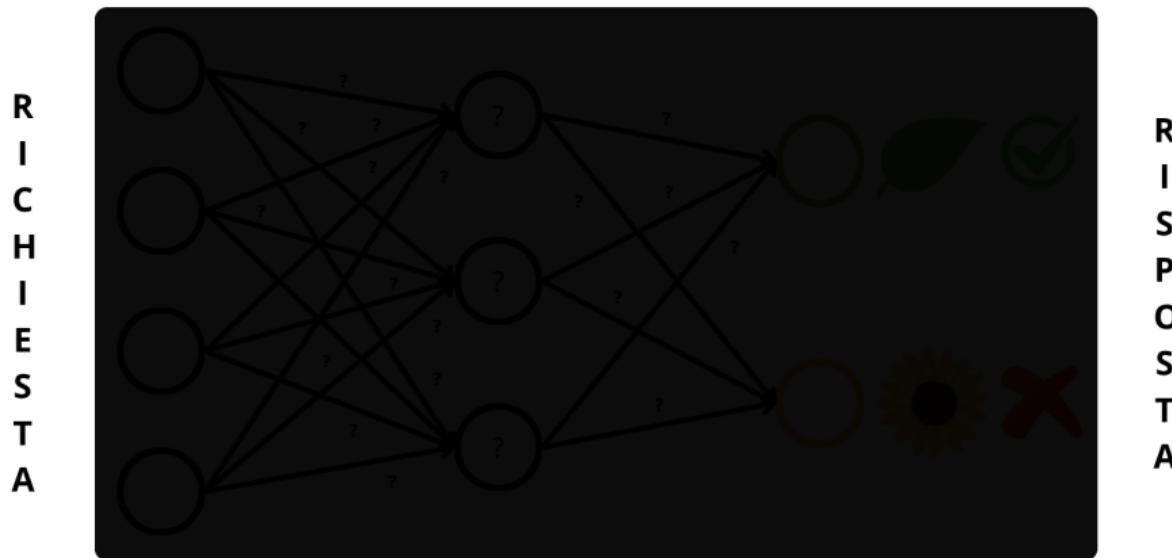


Figura 5: creata con Canva

PROBLEMI

PROBLEMI DELL'INTELLIGENZA ARTIFICIALE GENERATIVA

- **BIAS NEI DATI DI ADDESTRAMENTO:** pregiudizi e discriminazioni;
- **ALLUCINAZIONI:** informazioni false o fuorvianti;
- **MANIPOLAZIONE E DISINFORMAZIONE:** creazione di contenuti verosimili e/o ingannevoli;
- **PRIVACY E SICUREZZA INFORMATICA:** esposizione di dati sensibili;
- **ETICHETTATORI E SFRUTTAMENTO LAVORATIVO:** sfruttamento lavorativo e emotivo;
- **DIRITTO D'AUTORE E PROPRIETÀ INTELLETTUALE:** violazione del copyright;
- **SALUTE MENTALE E EFFETTI PSICOLOGICI:** effetti sulla salute mentale;
- **IMPATTO AMBIENTALE E SOSTENIBILITÀ:** consumo energetico e inquinamento;
- **ARMI AUTONOME:** decisioni cruciali autonome.

BIAS NEI DATI DI ADDESTRAMENTO

DEFINIZIONE

Il **bias AI**, chiamato anche bias del machine learning o bias dell'algoritmo, si riferisce al verificarsi di risultati distorti a causa di pregiudizi umani presenti nei **dati di addestramento** originali o l'algoritmo AI, portando a output distorti e potenzialmente dannosi.

[Appronfondimento](#)

[Esempi e rischi concreti](#)

BIAS NEI DATI DI ADDESTRAMENTO

ALGORITMI

All'intelligenza artificiale di Amazon non piacciono le donne: scartati i cv femminili

ANTI-FRODE ED ETICA DELL'AI

L'algoritmo dei sussidi sociali discrimina e fa cadere il governo: il caso olandese

L'intelligenza artificiale non è neutrale: può imparare i pregiudizi della nostra società

Vent'anni fa, *Minority Report* aveva già capito perché la polizia predittiva avrebbe fallito

Riconoscimento facciale. Secondo uno studio la polizia del Regno Unito non rispetta gli standard minimi etici e legali.

Figura 6: Immagine creata utilizzando screenshots tratti dai seguenti articoli: CorCom, Wired, Agenda Digitale, IRPA, Geopop

ALLUCINAZIONI

DEFINIZIONE

L'**allucinazione** (detta anche confabulazione) è una risposta generata da un'intelligenza artificiale generativa che contiene dati falsi o fuorvianti presentati come fatti **errati ma verosimili**. Il termine deriva da una vaga analogia con le allucinazioni umane, derivate normalmente da false percezioni. Le allucinazioni dell'intelligenza artificiale tuttavia, al contrario di quelle umane, derivano da risposte costruite in modo errato, e non da un'errata percezione.

Appronfondimento

[Dichiarazioni di OpenAI sulle allucinazioni](#)

ALLUCINAZIONI

The screenshot shows a news article from **Il Disinformatico** with the following content:

Podcast RSI - Story: Perché le Tesla vedono i fantasmi?

Colla sulla pizza: Google rimuove manualmente le risposte imbarazzanti di AI Overview

ChatGpt ha un problema di diffamazione

Ha accusato un professore di diritto di un'università statunitense di molestie e un sindaco australiano di corruzione. Ma in entrambi i casi, si è inventato di sana pianta tutto

McDonald's abbandona l'intelligenza artificiale dopo gli errori comici, dal bacon sul gelato alle 260 scatole di Nuggets

At the top right, there is a ChatGPT 4.0 interface showing the question: "How many R's are in the word strawberry" and the AI's response: "There are two 'R's in the word 'strawberry.'".

Figura 7: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [Il Disinformatico](#), [16x Prompt](#), [Wired](#), [Corriere della Sera](#), [Hardware Upgrade](#)

MANIPOLAZIONE E DISINFORMAZIONE

MANIPOLAZIONE E DISINFORMAZIONE

Sora 2, la nuova app di OpenAI è invasa da inquietanti video con celebrità morte

Podcast RSI – Arriva Sora, il ChatGPT dei video, ed è caos

Gli assistenti IA sono imprecisi nel dare le notizie: il 45% delle risposte contiene errori gravi. Lo studio della BBC

L'intelligenza artificiale generativa segna l'inizio di una nuova era per la disinformazione

Figura 8: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [Wired](#), [Il Disinformatico](#), [Dday.it](#), [Facta](#)

REALITY CHECK

Quiz: reale o generata artificialmente?

ENSHITTIFICATION

Le piattaforme online amplificano l'**AI slop** ovvero contenuti digitali realizzati con l'intelligenza artificiale generativa, in particolare quando vengono percepiti come privi di impegno, qualità e significati profondi e caratterizzati da un volume di produzione eccessivo.

MANIPOLAZIONE E DISINFORMAZIONE

“Vedere che il lascito di persone reali viene condensato fino a diventare un ‘questo somiglia vagamente e parla vagamente come loro ed è sufficiente così’, in modo che altre persone possano spandere orribile sbobba per TikTok che li muove come marionette, mi fa infuriare. [...] Non state facendo arte, state fabbricando hotdog disgustosi e ultraprocessati usando le vite di esseri umani, la storia dell’arte e della musica, e poi li cacciate in gola a qualcun altro sperando che vi diano un pollice alzato e a loro piacciono. Che schifo. [...] Smettete di chiarmarlo ‘il futuro’: l’intelligenza artificiale non fa che riciclare e rigurgitare il passato per consumarlo di nuovo. State ingerendo lo Human Centipede dei contenuti, e lo fate dal fondo della fila, mentre quelli in testa ridono, consumano e consumano.”

Citazione

Zelda Williams, figlia di Robin Williams

PRIVACY E SICUREZZA INFORMATICA

DEFINIZIONE

Ogni **input** inserito in un chatbot può diventare parte di un'enorme **banca dati**, usata per addestrare altri modelli o **esposta**, in casi estremi, a fughe di dati o accessi non autorizzati. Ecco perché è essenziale porsi delle domande:
cosa succede a ciò che scriviamo? Chi può leggerlo? Come possiamo proteggerci?

Appronfondimento

[Strategie pratiche per la privacy nei chatbot](#)

Stanchi delle risposte filtrate e limitate di ChatGPT? Ecco la versione 'jailbreak' DAN

Riconoscimento facciale: il Garante privacy sanziona Clearview per 20 milioni di euro. Vietato l'uso dei dati biometrici e il monitoraggio degli italiani

Segnalazione a Character AI

Il clamoroso fallimento di Builder, la startup che invece di un'AI aveva 700 programmati indiani

Cina, una scuola misura l'attenzione degli studenti con il riconoscimento facciale

Meta plans to sell targeted ads based on data in your AI chats

Il docente riceve in tempo reale i risultati del monitoraggio delle espressioni degli studenti, per capire se sono attenti e se apprezzano la lezione: l'ultima frontiera del Grande Fratello cinese

Figura 9: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [Hardware Upgrade](#), [Privacy Network](#), [Wired](#), [Garante per la protezione dei dati personali](#), [La Repubblica](#)

ETICHETTATORI E SFRUTTAMENTO LAVORATIVO

DEFINIZIONE

L'**etichettatura** dei dati richiede l'identificazione dei dati non elaborati (ad esempio immagini, file di testo, video) e quindi l'**aggiunta di una o più etichette a tali dati per specificarne il contesto** per i modelli, consentendo al modello di machine learning di fare previsioni accurate.

Appronfondimento

Data Labeling

Why Big Tech pays poor Kenyans to teach self-driving cars

Captcha if you can: how you've been training AI for years without realising it

Il lato oscuro di ChatGPT: lavoratori pagati meno di 2 dollari l'ora ed esposti alle peggiori nefandezze di Internet

Dietro ChatGPT c'è un esercito di addestratori sottopagati

Un'inchiesta di *Time* ha svelato che OpenAI ha affidato a un fornitore esterno il compito di visionare ed etichettare contenuti violenti per ripulire i risultati dell'algoritmo

Gli «etichettatori», ovvero i nuovi schiavi dell'intelligenza artificiale

Figura 10: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [BBC](#), [Forbes](#), [Wired](#), [TechCrunch](#), [TechRadar](#), [Dday.it](#)

DIRITTO D'AUTORE E PROPRIETÀ INTELLETTUALE

DEFINIZIONE

Il **diritto d'autore** è una branca del diritto privato, che ha lo scopo di **tutelare i frutti dell'attività intellettuale di carattere creativo** (ovvero le opere devono essere nuove e originali), attraverso il riconoscimento all'autore originario (o agli autori in caso di collaborazione creativa) dell'opera di una serie di diritti di carattere sia morale, sia patrimoniale.

Appronfondimento

[Il diritto d'autore](#)

Le AI hanno un grosso problema con i dati e con i copyright: cosa sta succedendo

Meta sotto accusa: avrebbe usato contenuti pirata ottenuti via Torrent per allenare l'IA

DIRITTO D'AUTORE

AI e copyright, la soluzione è nelle licenze esclusive

INTELLIGENZA ARTIFICIALE

AI, primo accordo sui diritti musicali: quali impatti sul settore

Figura 11: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [Geopop](#), [Hardware Upgrade](#), [Agenda Digitale: licenze esclusive](#), [Agenda Digitale: diritti musicali](#)

SALUTE MENTALE E EFFETTI PSICOLOGICI

SALUTE MENTALE E EFFETTI PSICOLOGICI

La nuova economia della solitudine

Il bisogno di ascolto in una società sempre più sola e frammentata può essere colmato dalle AI Personas?

Un Gesù virtuale sta confessando i fedeli in Svizzera

Innamorato di un chatbot, un adolescente americano si è suicidato a 14 anni. La madre fa causa all'app: «Pericolosa e non testata»

Dipendenza psicologica dall'intelligenza artificiale: un fenomeno da non sottovalutare

INTELLIGENZA ARTIFICIALE

L'ascesa dei virtual influencer: cosa ci dicono sul concetto di umanità

Riapre Way of the Future, la chiesa che venera gli algoritmi

DIGITAL AFTERLIFE: CONTINUARE A VIVERE OLTRE LA MORTE

Figura 12: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [Siamomine](#), [Agenda Digitale](#), [Wired](#), [Corriere della Sera](#), [Multiplayer.it](#), [Ansa](#), [Digital Innovation Days](#)

IMPATTO AMBIENTALE E SOSTENIBILITÀ

IMPATTO AMBIENTALE E SOSTENIBILITÀ

INTELLIGENZA ARTIFICIALE

L'intelligenza artificiale è pronta a favorire un aumento del 160% della domanda di energia dei data center

INDUSTRIA

Una nuova collaborazione tra Kairos, TVA e Google

Mar, 19 agosto 2025, 18:56 | Notizie Nucleari

*Podcast RSI - L'IA ha troppa fame di energia. Come metterla a dieta

INTELLIGENZA ARTIFICIALE

La sfida dei data center:
alimentare l'AI senza distruggere il pianeta

I Costi Planetari dell'Intelligenza Artificiale

CLIMATE CHANGE AND ENERGY

We did the math on AI's energy footprint. Here's the story you haven't heard.

The emissions from individual AI text, image, and video queries seem small—until you add up what the industry isn't tracking and consider where it's heading next.

Figura 13: Immagine creata utilizzando screenshots tratti dai seguenti articoli: [Goldman Sachs](#), [Il Disinformatico](#), [Nuclear Newswire](#), [InfoAut](#), [MIT Technology Review](#), [Agenda Digitale](#)

ARMI AUTONOME



Figura 14: Immagine di copertina del dossier di Valori.it: L'intelligenza artificiale va al fronte