

# *INTELLIGENZA ARTIFICIALE DA TURING A CHAT-GPT*

Angelo Montanari

Dipartimento di Scienze Matematiche, Informatiche e Fisiche

Università degli Studi di Udine

Polcenigo, 14 maggio, 2024

# L'Intelligenza Artificiale è nello spirito del tempo (zeitgeist)

Intervento di Toby Walsh (NICTA, University of New South Wales, Australia), in una delle principali conferenze internazionali di IA (KR) nel 2016: **Will AI end jobs, wars or humanity?** Cape Town, SA, April 2016:

- The Chief Economist of the Bank of England some years ago predicted **AI will destroy 50% of jobs in the UK.**
- Thousands of AI researchers signed an Open Letter predicting that AI could transform warfare and lead to an arms race of **“killer robots”**.
- S. Hawking and others have predicted that **AI could end humanity itself.**

Nel 2023, poco dopo l'esplosione di Chat GPT, venne pubblicato un appello che chiedeva: **“Immediately suspend training of systems more powerful than gpt-4 for at least 6 months”**, Future of Life Institute (E. Musk, S. Wozniak, Y. Bengio e molti altri).

# Né idolatrare né demonizzare

L'uomo costruisce gli **idoli** e poi li venera (Esodo 32), ma gli idoli (come le macchine, per quanto sofisticate) sono prodotti dell'essere umano.

Il problema può essere dimenticare l'origine dei sistemi artificiali intelligenti e idolatrarli (o demonizzarli).

Ciò non vuol dire che il **rapporto con tali macchine** sia facile:

- da sempre le macchine fanno delle cose che l'uomo non è in grado di fare (pensiamo alla rivoluzione industriale e ai mezzi di trasporto);
- il problema è sostituire e, spesso, migliorare le prestazioni dell'essere umano in compiti ritenuti da sempre di sua competenza esclusiva (**intelligenza artificiale: ragionamento e memoria**).

# Un vocabolario antropomorfo

- L'uso di un vocabolario antropomorfo nella descrizione delle caratteristiche e del funzionamento dei sistemi informatici
  - è particolarmente evidente nel caso dei sistemi di intelligenza artificiale (intelligenza, conoscenza, apprendimento, ragionamento),
  - ma si è verificato in misura più o meno rilevante in molti altri ambiti dell'informatica (memoria, comunicazione, interrogazione).
- Ragioni dell'uso di un vocabolario antropomorfo
  - Uomo/animale come modello in cibernetica e successivamente in diversi settori dell'informatica (Intelligenza Artificiale, Robotica, Bionica, ..).

# Norbert Wiener

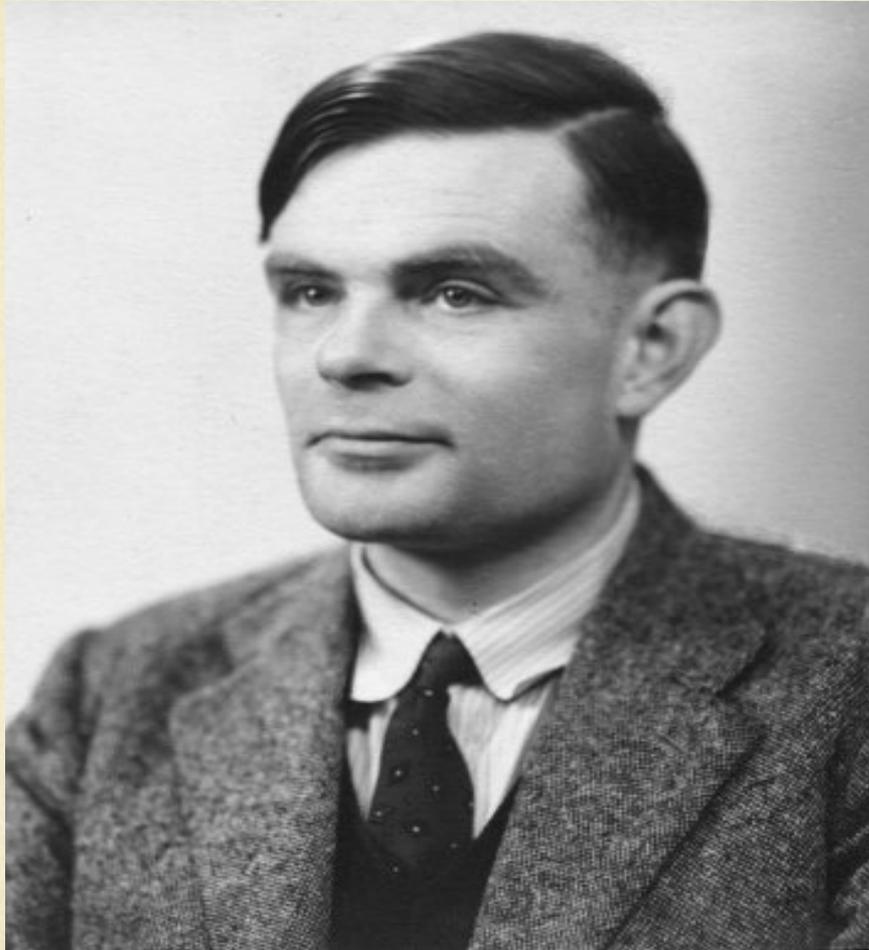


- Norbert Wiener: *Cybernetics or Control and Communication in the Animal and the Machine*, MIT Press, 1962).

# Che cos'è l'Intelligenza Artificiale?

- Per quanto rilevante, l'**Intelligenza Artificiale** è da sempre un importante settore dell'**informatica** (era già presente negli scritti di Alan Turing, il padre riconosciuto dell'informatica).
- Lo specifico dell'informatica: la capacità di **modellare** e di **risolvere problemi**, dove la prima abilità è almeno tanto importante quanto la seconda.
- Il concetto chiave di **algoritmo**: una descrizione finita e non ambigua di una sequenza di operazioni che consente ad un certo agente di risolvere un determinato **problema**.

# Alan Turing



Computing Machinery and Intelligence, A. M. Turing, *Mind*, New Series, Vol. 59(236), October 1950, pp. 433-460.

# Il test di Turing

Il **test di Turing** (o gioco dell'imitazione): una macchina può essere definita intelligente se riesce a convincere una persona che il suo comportamento, dal punto di vista intellettuale, non è diverso da quello di un essere umano medio.



# Intelligenza e linguaggio

Il Test di Turing assume un legame molto stretto tra intelligenza e **linguaggio** (naturale): l'intelligenza si manifesta nell'interazione/comunicazione attraverso il linguaggio.

Il linguaggio occupa un ruolo fondamentale nel rudimentale **sistema esperto ELIZA** (psicanalista digitale) proposto da Weizenbaum negli anni '60.

La comprensione di testi in linguaggio naturale è al centro del famoso **esperimento mentale della stanza cinese** di Searle (impossibilità per una macchina di manifestare l'**intenzionalità** che caratterizza gli esseri umani e, sia pure in forme diverse, gli animali).

All'ambito dell'elaborazione del linguaggio naturale appartiene anche l'universo dei sistemi conversazionali (Chatbot), dei quali **ChatGPT** è uno dei rappresentanti più noti.

# Intelligenza e corporeità

Il modello di intelligenza sotteso al Test di Turing è un modello astratto/disincarnato dell'intelligenza.

Una delle acquisizioni più importanti della ricerca in IA degli ultimi decenni è la consapevolezza del ruolo cruciale che gli organi di senso svolgono nell'interazione dell'uomo col mondo e della conseguente impossibilità di un'intelligenza (artificiale) priva di "corporeità". Ciò ha portato allo sviluppo di un rapporto sempre più stretto tra IA ("cervello senza corpo") e robotica ("corpo senza cervello").

Per paradossale che possa suonare, per avvicinarsi all'intelligenza umana l'IA deve diventare un'intelligenza incarnata.

# I robot umanoidi

A partire dal 2009, il Centro Aerospaziale Tedesco ha sviluppato una famiglia di robot umanoidi programmabili (Justin e le sue varianti) in grado di operare con una significativa autonomia.



# La robotica collaborativa

La robotica collaborativa combina le potenzialità e le capacità delle persone e dei robot per portare a termine, in maniera semi-automatica, dei compiti che risultava impossibile realizzare in maniera completamente automatica.

Che cosa sono i robot collaborativi (**cobot**)?

Sono dei robot industriali (ma vi sono applicazioni anche nell'ambito dell'assistenza domiciliare e della domotica) di nuova generazione progettati per lavorare assieme all'essere umano, fianco a fianco e in sicurezza, senza la necessità di barriere o gabbie protettive che separino gli uni dagli altri.

# IA guidata dai modelli e IA guidata dai dati

In IA sono da sempre presenti **due filoni** principali: l'IA guidata dai modelli (IA simbolica) e IA guidata dai dati (IA sub-simbolica).

**IA simbolica:** rappresentazione della conoscenza e ragionamento automatico. Un caso paradigmatico: la pianificazione automatica.

**IA sub-simbolica:** mimare il comportamento del cervello umano e la sua complessa rete di neuroni interconnessi. I primi modelli di rete neurale risalgono agli anni '40 (il modello dei neuroni di McCulloch e Pitts, 1943).

Il caso dell'**elaborazione del linguaggio naturale:** dalla logica ai corpora (collezioni di grandi dimensioni di testi orali o scritti prodotti in contesti comunicativi reali).

# L'esplosione del machine/deep learning

Quali sono le ragioni dell'esplosione della ricerca e delle applicazioni del machine/deep learning?

Enorme **potenza di calcolo** ed enorme disponibilità di **dati** (i dati sono il nuovo **petrolio**).

Il caso di **ChatGPT (Generative Pre-trained Transformer)**: chatbot in grado di dialogare con l'utente in linguaggio naturale

- 1.800.000.000.000 parametri

- dataset di training di 1.000 Terabyte (1 Terabyte = 1.000 Gigabyte; 1 Gigabyte = 1 miliardo di byte).

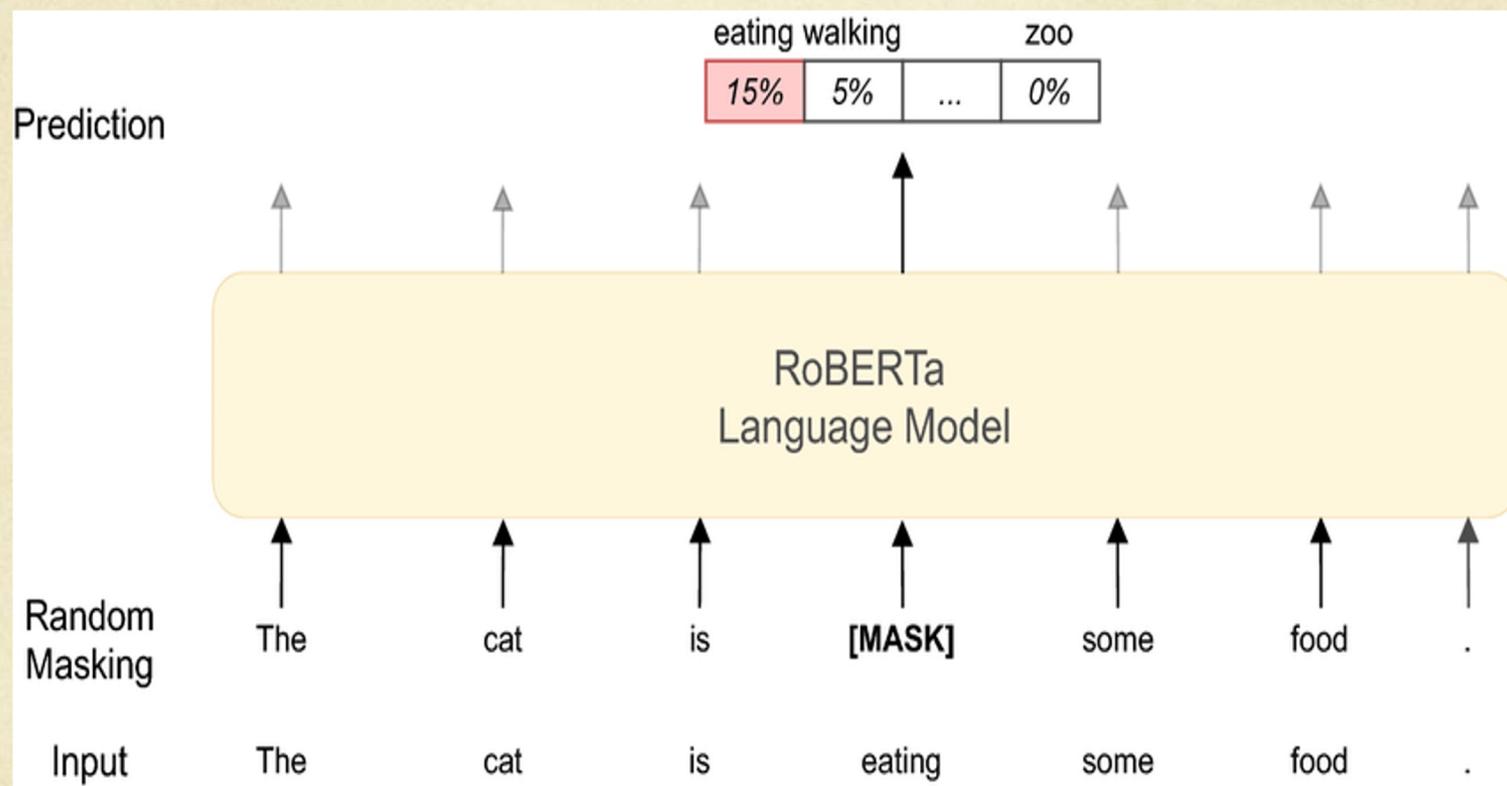
# ChatGPT: modelli del linguaggio

Un **modello del linguaggio** (LLM - Large Language Model) cerca di catturare la struttura e le regole del linguaggio naturale (ad esempio, della lingua italiana).

Per addestrare un modello del linguaggio, si utilizza un insieme molto grande di testi (corpus) contenente un'ampia varietà di frasi in linguaggio naturale. Tramite l'analisi di tale corpus, l'algoritmo cerca di indentificare strutture e regole grammaticali e, soprattutto, relazioni «statistiche» fra le parole.

Possibile tecnica: data una frase, si «maschera» una parola e si insegna alla rete neurale a ricostruire tale parola sulla base di quelle che la circondano (contesto).

# Un esempio: RoBERTa Language Model



# ChatGPT: come raffinare un modello?

Dato un modello del linguaggio, occorre specializzarlo affinché sia in grado di sostenere una conversazione (chat) con un essere umano nel modo più realistico possibile.

A tal fine, viene esaminato un gran numero di conversazioni fra esseri umani su svariati argomenti. Il modello cerca di capire quali sono i tratti distintivi di ogni conversazione e tenta di replicarli, in modo da consentire al chatbot di rispondere in modo appropriato (e naturale).

Esempio. Alla domanda: «Come stai?», il modello potrebbe rispondere: «Bene, grazie. E tu?».

Possono essere utilizzati anche tecniche e strumenti di **reinforcement learning**.

# Il meccanismo fondamentale

Risultato: un chatbot che risponde basandosi su regole statistiche, cercando la risposta più realistica possibile.

Meccanismo fondamentale: data una sequenza di parole/caratteri (di lunghezza  $n$ ), stabilire quale nuova/o parola/carattere generare (ottenendo una sequenza di lunghezza  $n+1$ ).

ChatGPT viene usato per svolgere diversi compiti, inclusi prompt engineering, generazione automatica di riassunti, estrazione di informazioni, analisi dei sentimenti, individuazione degli argomenti presenti in un testo, traduzione, generazione automatica di risposte a messaggi di clienti, produzione di immagini o di video.

# Alcuni esempi (1)

A

Potresti aiutarmi a progettare la mia pagina web?



Certo! Come posso aiutarti a progettare la tua pagina web? Hai già un'idea di cosa vuoi includere sulla tua pagina web?

A

Se Mario ha 3 mele, ne mangia 1, ma poi sua zia gliene regala 2, quante mele ha Mario alla fine?



Dopo aver mangiato una mela, Mario ne ha rimaste 2. Quando sua zia gliene regala 2, Mario ha in totale 4 mele. Quindi alla fine Mario ha 4 mele.

# Alcuni esempi (2)

A

Potresti scrivere un testo horror di 100 parole su un toast incastrato in un videoregistratore?



Il toast sembrava normale, ma quando l'ho inserito nel vecchio videoregistratore, qualcosa di strano è successo. Lo schermo si è acceso da solo e ho visto una figura spettrale che si materializzava lentamente. Ho cercato di estrarre il toast, ma era incastrato, come se fosse stato risucchiato dentro il videoregistratore. La figura spettrale si avvicinava sempre di più, finché non ho sentito la sua presenza fredda e inquietante sulla mia pelle. Mi sono svegliato in un grido, ma il toast è ancora intrappolato, come se volesse trascinarci di nuovo nell'incubo.

# L'altra faccia della medaglia

L'uso di strumenti come ChatGPT ha ovviamente dei costi:

- costi economici;
- costi ambientali (consumo di energia / impatto ecologico);
- costi sociali (di diversa natura e più difficili da misurare).

Un esempio: la ricerca e lo sviluppo in tale ambito è svolta sempre meno in ambito pubblico/accademico e sempre più in ambito privato/aziendale.

# Trustworthy AI

**Human-in-the-loop:** un modello/sistema che prevede l'interazione col soggetto umano (ad esempio, nei processi di decisione).

Non sempre è possibile garantire tale condizione (applicazioni finanziarie e sistemi critici dal punto di vista della sicurezza).

In alcuni casi, è essenziale (diagnosi mediche «critiche»).

Obiettivo: **trustworthy AI** (intelligenza artificiale affidabile).

L'IA di cui ti puoi fidare, perché è in grado di rendere ragione/spiegare le conclusioni/previsioni formulate.

L'**AI Act** europeo: come coniugare libertà e responsabilità nell'IA (approvato il 14 marzo 2024 dal Parlamento Europeo).

# A che punto siamo?

L'obiettivo trustworthy AI non è (ancora) stato raggiunto.

«ChatGPT è una promessa che deve essere accolta con estrema attenzione/prudenza» D. Pedreschi, Partenariato esteso FAIR (Future Artificial Intelligence Research).

ChatGPT non sa che cosa non sa, ossia non è in grado di dare una misura della robustezza/affidabilità delle sue risposte.

Il problema delle allucinazioni.

ChatGPT sta usando le persone, spesso a loro insaputa, per la validazione e l'affinamento delle sue capacità («fine tuning»).

Come riconoscere i testi generati dai Large Language Model?

# Un esempio di argomentazione fallace



A dog run around a square field at constant speed. It takes it 2min per side, except the last one for which it takes it only 120s, why?



If the dog runs around a square field at a constant speed, it should take the same amount of time to run each side of the square. However, the problem states that it takes the dog only 120 seconds to run the last side of the square, while it takes 2 minutes (or 120 seconds) to run each of the other three sides.

One possible explanation for this is that the dog is able to accelerate during the final side of the square. This could be due to several reasons, such as the terrain being more favorable for the dog to run faster or the dog being motivated to finish the last side quickly.

Another possible explanation is that there is a measurement error in the problem. For example, it is possible that the time taken to run the last side was not accurately measured, or that there was a mistake in recording the time taken to run the other sides.

Without further information, it is difficult to say for certain why the dog took only 120 seconds to run the last side of the square.



# Come porsi nei confronti dell'IA?

La valutazione di un ricercatore di IA non può che essere **positiva**.

Un **esempio** fra i tanti possibili (l'IA è ovunque, dalla guida autonoma ai chatbot): **early failure detection** e **predictive maintenance** (dagli impianti industriali e i sensori alla salute delle persone).

Non possono essere dimenticate tutte le **criticità** di diversa natura (psicologica, pedagogica, sociale, culturale, politica).

I **dati fotografano il passato**: il comportamento umano ripetitivo (dalle vacanze alla politica) è prevedibile/replicabile e, di conseguenza, influenzabile.

Elogio della **creatività**: dobbiamo esercitare il pensiero critico ed imparare ad esplorare nuove strade (futuro aperto).