Coscienza artificiale (Intelligenza Artificiale e coscienza delle macchine)

Angelo Montanari
Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Università degli Studi di Udine

San Vito al Tagliamento, 4 luglio, 2023

L'Intelligenza Artificiale (IA) è nello Zeitgeist (spirito del tempo)

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:

- Il capo economista della Banca d'Inghilterra alcuni anni fa ha previsto che l'**IA distruggerà il 50% dei posti di lavoro nel Regno Unito** (la scorsa settimana, Graziano Tilatti, presidente di Confartigianato: «**106.000 lavoratori ad alto rischio in Fvg**»).

- Migliaia di ricercatori di IA hanno firmato una lettera aperta prevedendo che l'IA potrebbe **trasformare la guerra** e portare a una corsa agli armamenti di «**robot assassini**» (l'uso dei **droni** nella guerra in corso in **Ucraina**).

- S. Hawking e altri hanno previsto che l'IA potrebbe **porre fine all'umanità** stessa.

Un paio di mesi fa, il Future of Life Institute (E. Musk, S. Wozniak, Y. Bengio e molti altri) ha chiesto di «sospendere immediatamente l'addestramento di sistemi più potenti di GPT-4 per almeno 6 mesi».

Intelligenza Artificiale: entusiasti e apocalittici

E' possibile stendere un elenco di **applicazioni** presenti e future dell'IA che copre tutte le lettere dell'alfabeto, dalle **a**uto senza pilota alla **b**ionica, da **C**hat-GPT alla **d**omotica, fino alla **z**, passando, ad esempio, per la **r**obotica collaborativa.

- **Entusiasti** (la filosofia postumanistica/transumanistica e il superamento dell'umano): Marvin Minsky (la società della mente), Raymond Kurzweil (la singolarità tecnologica).
- **Apocalittici**: Stephen Hawking, ma non solo.
- Si va verso nuove forme di **luddismo** (lotta violenta contro l'introduzione di nuove machine artificiali)?
- Un caso esemplare: la curiosa denigrazione dell'algoritmo.

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** non sia problematico:

- 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);
- la novità è che l'IA è in grado di sostituire l'essere umano e, spesso, di migliorarne le prestazioni in compiti ritenuti da sempre di sua competenza esclusiva (**intelligenza artificiale: ragionamento** e **memoria**).

Un vocabolario antropomorfico

- Un altro punto importante: l'uso di un **vocabolario antropomorfico** 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 tale vocabolario:
  - uomo/animale come modello di riferimento 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).

Il rapporto uomo/macchina

- Ogni **discorso** sulle proprietà "antropomorfiche" delle macchine/calcolatori non riguarda tanto la macchina (il calcolatore) in sé, ma il modo in cui noi vediamo la macchina, e indirettamente noi stessi.

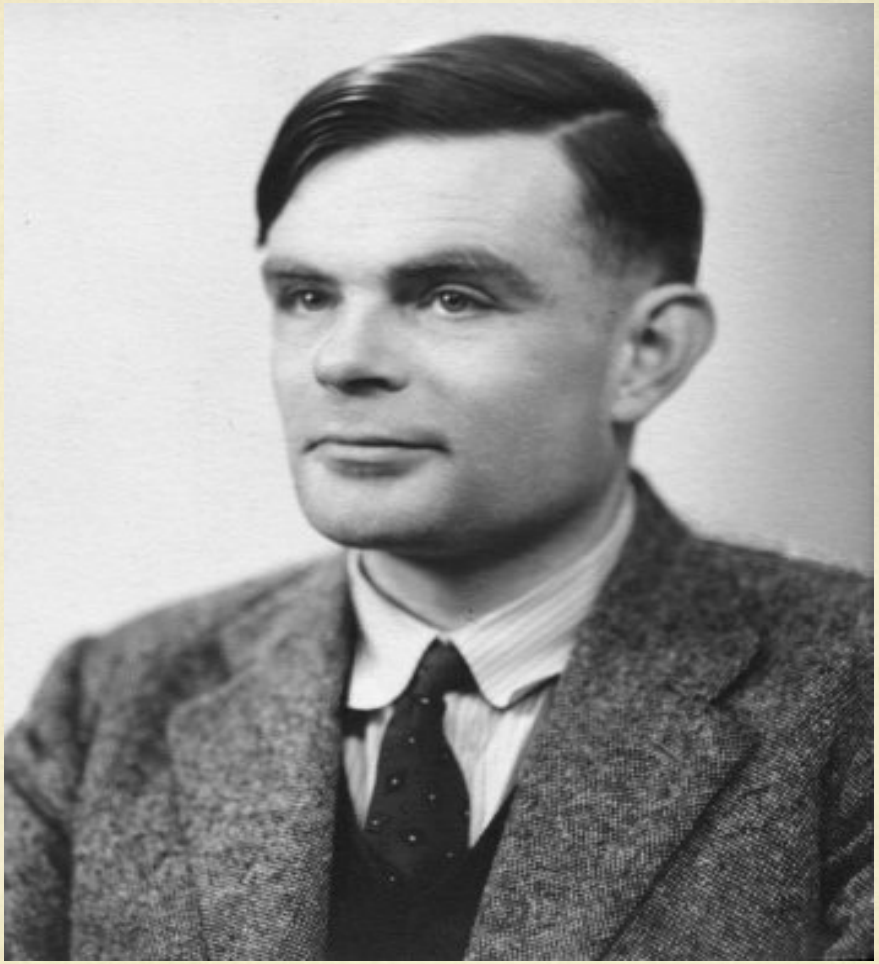
Ad esempio, Minsky rivendica la legittimità/utilità dell'uso di termini antropomorfici (per Minsky l'uomo è «una macchina meravigliosa»).

- Ciò vale, ad esempio, per la questione relativa alla **coscienza/intenzionalità** delle **macchine**: parlare delle macchine è un modo (indiretto) per parlare di noi stessi (è una questione antropologica).

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** (un esempio di algoritmo molto famigliare: una ricetta).

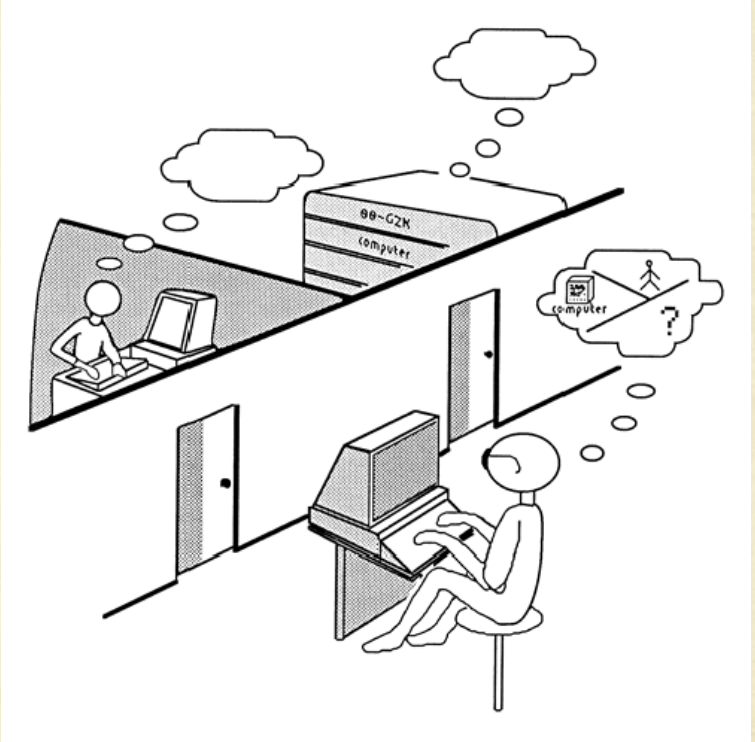
Alan Turing



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



IA guidata dai modelli o 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 (Chat-GPT)**: 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 (IA guidata dai dati)?

Enorme **potenza di calcolo** ed enorme disponibilità di **dati.**

Il caso di **ChatGPT (Generative Pre-trained Transformer)**: un miliardo di parametri.

L'altra faccia della medaglia:

- costi sociali;
- costi economici;
- costi ambientali (consumo di energia / impatto ecologico)

Trustworthy AI

Human-in-the-loop: un modello/sistema che prevade 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**. 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 (14 giugno 2023).

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, Coordinatore Dottorato Nazionale in IA e Partenariato esteso PNRR 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.

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

Coscienza artificiale?

Una questione antropologica.

Per poter ragionare di coscienza delle macchine (coscienza artificiale) occorre preliminarmente stabilire **cos'è la coscienza (naturale)**.

Posizioni fortemente divergenti.

Marvin Minsky: la coscienza ha più a che fare con le cose semplici/banali che non con le cose complicate/profonde.

John Searle: l'intenzionalità come caratteristica distintiva della coscienza e l'impossibilità per un sistema di IA, come quelli sin qui sviluppati, di possederla.

La coscienza secondo Minsky

La coscienza interviene quando qualcuno dei processi più semplici non funziona correttamente.

Nella mente di ogni persona sembrano esservi dei processi che chiamiamo **coscienza** e riteniamo che essi ci consentano di sapere che cosa accade nella nostra mente (la nozione di autoconsapevolezza).

In realtà, i nostri pensieri coscienti ci rivelano pochissimo di ciò che li genera. Essi inviano semplicemente dei segni/segnali che consentono di pilotare il motore della nostra mente, controllando innumerevoli processi di cui non siamo mai molto consapevoli.

Marvin Minsky

The Society of Mind,
TOUCHSTONE BOOK 1988
(tr. it. La società della mente,
Biblioteca Scientifica 10,
Adelphi, 1989).

La stanza cinese di Searle

Tratto distintivo della coscienza: l'**intenzionalità**.

Impossibilità per una macchina di manifestare l'intenzionalità che caratterizza gli esseri umani e, sia pure in forme diverse, gli animali.

Tesi: L'esecuzione di un algoritmo (programma) su un dato input non è mai di per se stessa una condizione sufficiente per l'intenzionalità.

La **dimostrazione** (un esperimento mentale): Searle immagina di sostituire un agente umano al calcolatore nel ruolo di esecutore di una specifica istanza di un programma e mostra come tale esecuzione possa avvenire senza forme significative di intenzionalità.

John Searle

John. Searle, *Minds, brains, and programs*, *Behavioral and Brain Sciences*, vol. 3, 1980.

Digressione: IA e creatività

- Non confondere l'incapacità di prevedere/controllare il comportamento di una macchina con un'attività creativa della macchina.
- Più diventa complessa/potente una macchina più risulta difficile controllarne il comportamento.
- Il problema dell'**autonomia** e della responsabilità delle macchine (si pensi alle auto a guida autonoma).

Come porsi nei confronti dell'IA?

La valutazione di un ricercatore di IA non può che essere **positiva**, pur nella consapevolezza di tutte le **criticità**.

Un **esempio** fra i tanti possibili: **early failure detection** e predictive maintenance (dagli impianti industriali e i sensori alla salute delle persone).

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

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

IA all'Università di Udine

Classifica «AI 2000» creata dalla Tsinghua University e dall'Associazione cinese di IA. Nel 2023, 4 docenti dell'Università di Udine sono stati inseriti fra i 2000 studiosi più influenti al mondo nella loro area di ricerca in IA. Sono Fabio Buttussi, Luca Chittaro, Angelo Montanari e Giuseppe Serra del Dipartimento di Scienze Matematiche, Informatiche e Fisiche.