My Fair AI

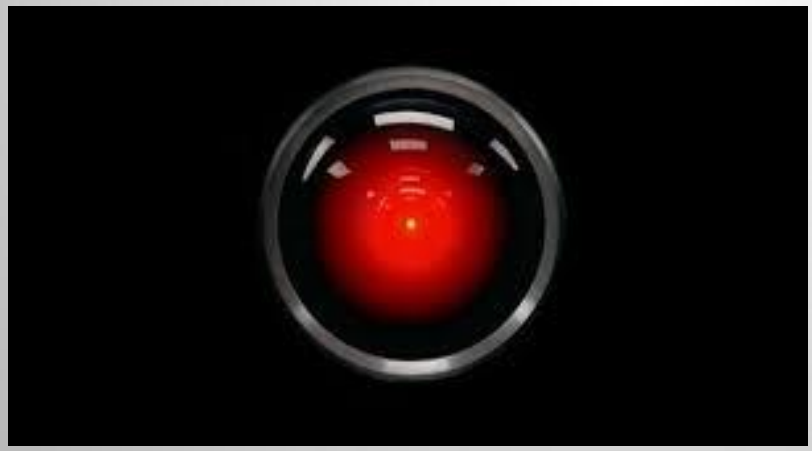
Tiziana Catarci

DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

L’obiettivo (il sogno?) dell’AI

- Da Turing (~1935/45) in poi:
  - Costruire un marchingegno intelligente
  - Un manufatto col quale si interagisce come se fosse un essere umano



Un po' di storia

Filosofia	'31 Gödel *Dimostrabilità*
Matematica	'43 Mc Cullogh Pitts *Reti Neurali*
Psicologia	'50 Turing *Machine Intelligence*
Economia	'56 McCarthy *Artificial Intelligence*
Linguistica	
Neuroscienza	'60 Primi Entusiasmi
Cibernetica	'70 Inverno AI
Informatica	'80 Sistemi Esperti
	'90 Agenti Intelligenti

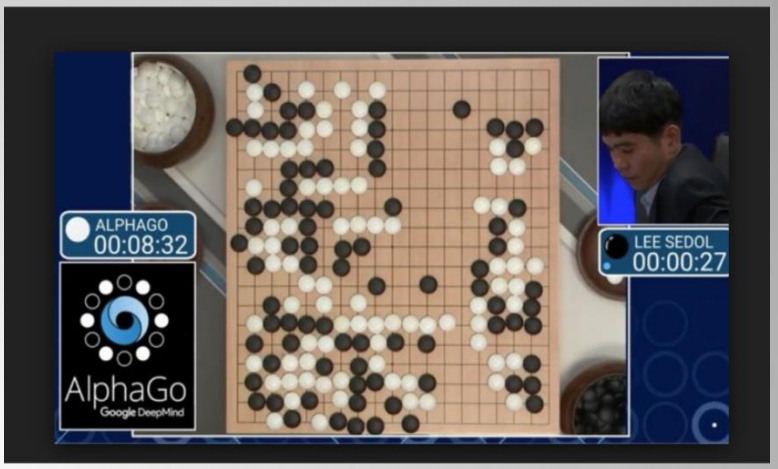
1992 Insegnamento AI@Sapienza

John McCarthy e il marketing nel nome «AI»

“Excuse me, I invented the term ‘Artificial Intelligence’. I invented it because we had to do something when we were trying to get money for a summer study in 1956, and I had a previous bad experience. The previous bad experience [concerns, McCarthy corrects himself and says] occurred in 1952, when Claude Shannon and I decided to collect a batch of studies, which we hoped would contribute to launching this field. And Shannon thought that ‘Artificial Intelligence’ was too flashy a term and might attract unfavorable notice. And so, we agreed to call it ‘Automata Studies’. And I was terribly disappointed when the papers we received were about automata, and very few of them had anything to do with the goal that at least I was interested in. So, I decided not to fly any false flags anymore but to say that this is a study aimed at the long-term goal of achieving human-level intelligence. Since that time, many people have quarrelled with the term but have ended up using it. Newell and Simon and the group at Carnegie Mellon University tried to use ‘Complex Information Processing’, which is certainly a very neutral term, but the trouble was that it didn't identify their field, because everyone would say ‘well, my information is complex, I don’t see what’s special about you’.» The Lighthill Debate (1973)

AI e le Sfide

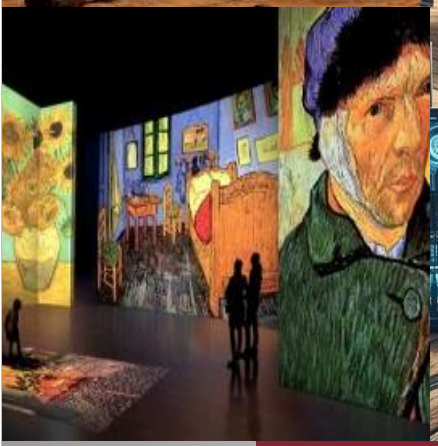
- 1997 Deep Blue (scacchi)
- 1997 RoboCup
- 2005 DARPA Challenge
- 2011 IBM Watson Jeopardy



- 2012 Deep Learning Imagenet
- 2016 Deepmind AlphaGO (GO)
- 2017-19 DeepStack/Pluribus win poker
- 2022 Large Language Models OpenAI ChatGPT

LA RIVOLUZIONE DIGITALE

DIGITAL IS EVERYWHERE



La Società Digitale

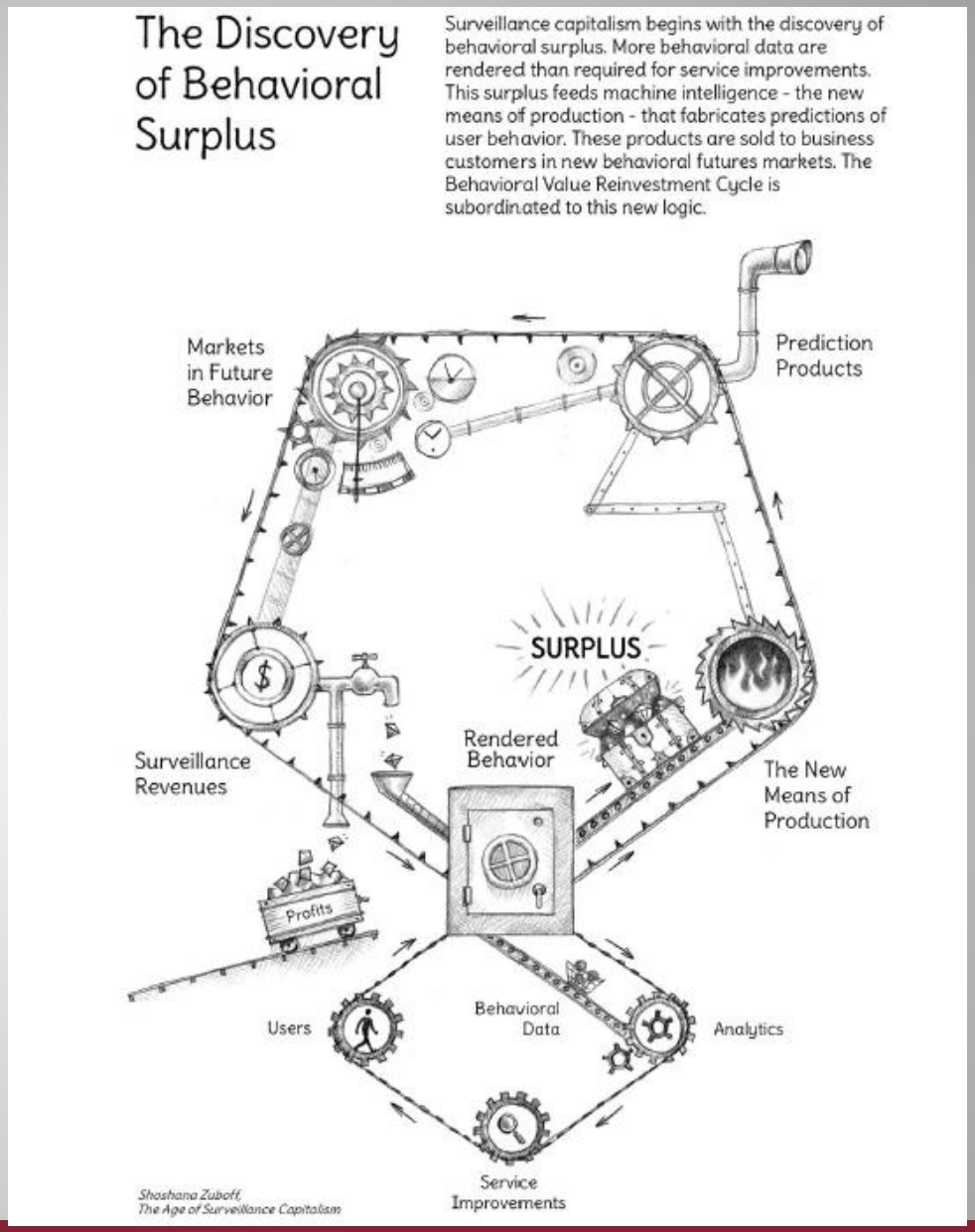
Nuova Rivoluzione Industriale, aumento smisurato dell'informazione disponibile, intelligenza artificiale, **grandi possibilità** e **grandi rischi**
Cambiamenti sociali, economici, culturali, urbanistici, politici,…
Nuovi potenti strumenti per risolvere i (vecchi e nuovi) grandi problemi del mondo

Cambiamenti nel mondo del lavoro, scomparsa dei lavori ripetitivi, emergere di nuovi lavori (creativi, innovativi). Nuovi lavori e competenze multi/interdisciplinari per affrontare problemi sempre più complessi: cambiamento climatico, energia, invecchiamento della popolazione, pandemie,…
Importanza del pensiero critico
Consapevolezza di possibilità e rischi
Alfabetizzazione Digitale
Competenze digitali per guidare il cambiamento e non subirlo
Valore della diversità

Amplificazione pregiudizi e disuguaglianze
Digital divide
Pensiero Unico
Disinformazione (anche »personalizzata»)
Polarizzazione (echo chambers e filter bubbles)
Consumo di Energia
Litio e altri metalli
Rapida obsolescenza tecnologica
Controllo degli esseri umani
Monopolizzazione del mercato (e della società)
Perdita di lavoro, nuovi lavori ripetitivi e sottopagati

L’inizio: le piattaforme e i dati

«New means of production/extraction (which generate behavioral surplus) enable companies to create “Prediction Products.” Those products let the company know what its users will do, how they will behave, and in turn let it trade in a *beharioval futures market*, where advertisers are willing to buy those future behaviors (clicks, sales, etc…). That enables the company to generate “*surveillance revenues*” (and profits), which it will reinvest in 2 areas: service improvements, to continue to grow the user base of the products, and more means of production/extraction, hence accelerating the cycle.»



Big Tech

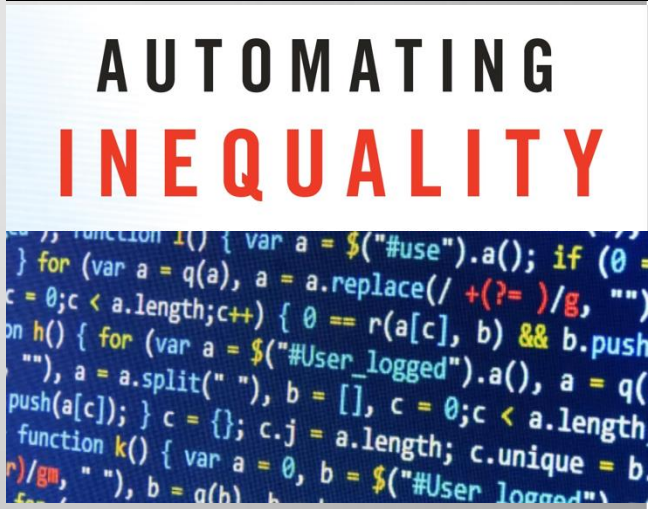
- Nel 2022 le big five - Alphabet (Google), Microsoft, Amazon, Apple, Meta - e Nvidia hanno investito complessivamente 210 miliardi di dollari in ricerca nel solo settore del digitale (cloud, social media, AI, sistemi mobili, metaverso, robotica e nuovi processori)
- Nel 2023, considerando anche Tesla (le «Magnificent Seven»), 309,2 miliardi di dollari.
- Meta (39 miliardi) investe nel solo digitale quanto l'Italia complessivamente in tutti i settori, Amazon (73 miliardi) quanto la Francia
- Come si può colmare questo divario senza (almeno) una strategia comune europea?

Training GPT3, circa 4.6M$ e 355 anni-GPU
Servizi di Chat-GPT circa 0.7M$ al giorno e 29.000 Nvidia A100 GPU

Un singolo algoritmo può influenzare le scelte di centinaia di milioni di persone e consentire la condivisione dei loro dati e delle loro informazioni in pochi istanti (quasi sempre 'gratuitamente').

Software Power

- La diffusione esponenziale delle tecnologie digitali sta fornendo **nuove soluzioni** a molti problemi.
- Il **digitale** si sta configurando come **una nuova forma di potere** che crea consenso.
- Una nuova forma di consenso di un numero enorme di individui che utilizzano strumenti e servizi automatici basati su algoritmi “intelligenti”.
- Mentre risolvono i nostri problemi, **gli algoritmi introducono nuove disuguaglianze** e, in alcuni casi, non aiutano i soggetti svantaggiati e deboli.



Warnings recenti

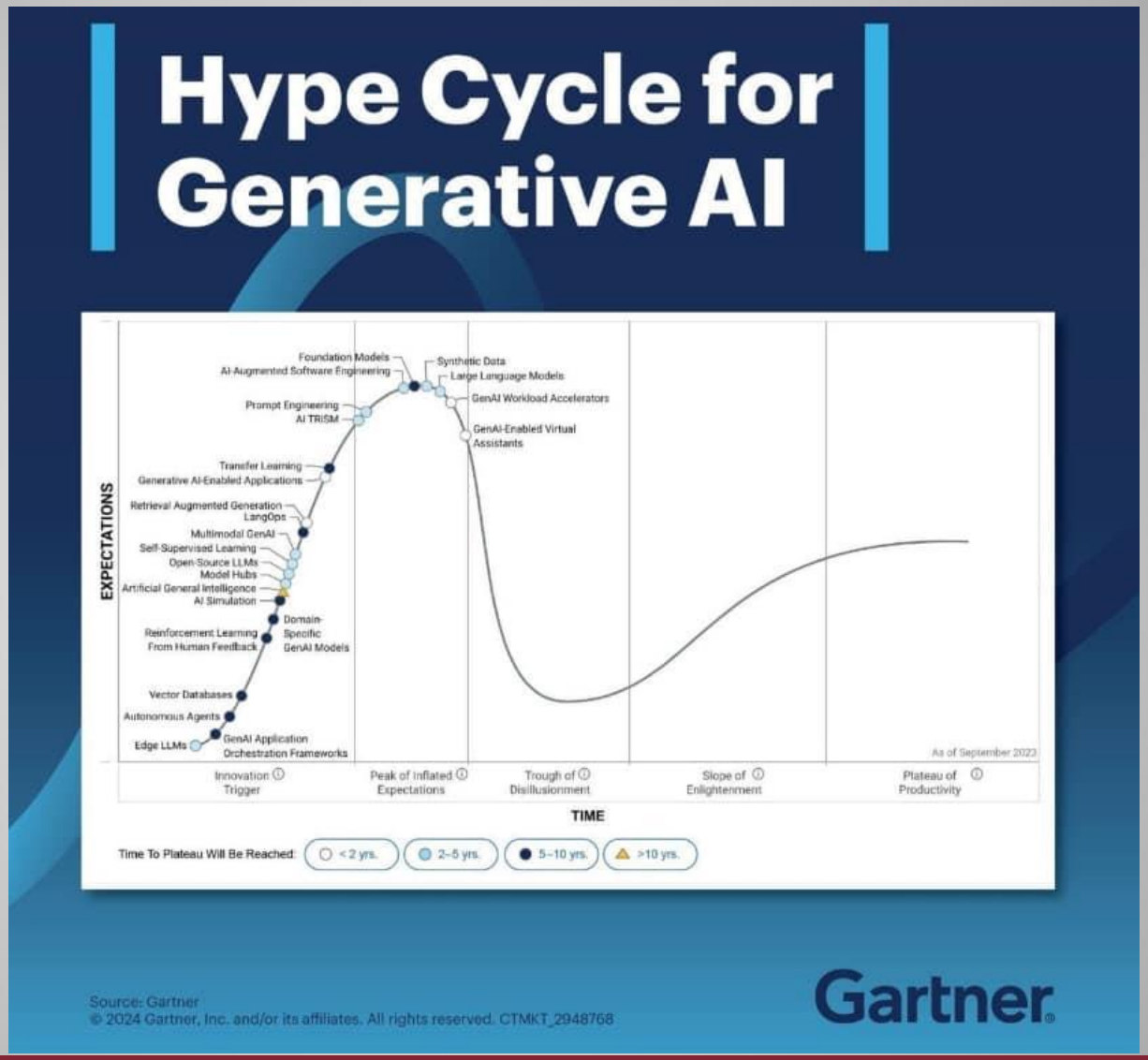
Digital Humanism Initiative open letter: «These threats emanate from an unprecedented confluence of immense political and economic power and the EU regulations may well constitute one of the last lines of defence against these threats worldwide…They have been designed to protect information and election integrity, protect from monopolistic power abuses and show that there are ways to harness the benefits of technology for all while mitigating its potential harms.»

Antiqua et Nova: «i dati finora raccolti sembrano suggerire che le tecnologie digitali siano servite ad aumentare le disuguaglianze nel mondo. Non solo le differenze di ricchezza materiale, che pure sono importanti, ma anche quelle di accesso all'influenza politica e sociale. In questo senso, l'IA potrebbe essere usata per protrarre situazioni di marginalizzazione e discriminazione, per creare nuove forme di povertà, per allargare il "divario digitale" e aggravare le disuguaglianze sociali…Inoltre, il fatto che attualmente la maggior parte del potere sulle principali applicazioni dell'IA sia concentrato nelle mani di poche potenti aziende solleva notevoli preoccupazioni etiche…Tali entità, motivate dai propri interessi, possiedono la capacità di esercitare forme di controllo tanto sottili quanto invasive, creando meccanismi di manipolazione delle coscienze e del processo democratico»

Quale AI?

- Generale e “forte” (orizzontale, applicabile a problemi diversi, vuole riprodurre tutte le funzionalità della mente umana) *
- Ristretta e “debole” (focalizzata su alcune funzionalità, verticale e applicabile a problemi specifici)
- Generativa (generalizza l’approccio usato per problemi specifici e genera nuovi output a partire dagli input)
- Intelligenza “aumentata”, essere umano + AI
- Symbolic AI (il mondo può essere concettualizzato come una struttura simbolica su cui ragionare e risolvere problemi)
- Connectionist AI (le decisioni si basano su una rete di nodi interconnessi, le cui interazioni crescono durante la vita del sistema)
- * Next step: Symbolic + Connectionist (eg neurosymbolic AI)? General AI

Previsioni per il Futuro



Definizione 2023 AI

OECD e AI-Act

An AI system is a machine-based system that can, for explicit or implicit objectives, infer, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment

Precedentemente: L'**intelligenza artificiale** è una disciplina che studia i fondamenti teorici, le metodologie e le tecniche che consentono la progettazione di sistemi hardware e software capaci di fornire prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana.

Esempio: Salute

- Migliorare la precisione delle diagnosi mediche
- Anticipare l’esito di possibili interventi terapeutici
- Focalizzare le campagne di prevenzione
- Far sviluppare la medicina personalizzata
- Definire campagne di cura e supporto (e selezionare le/i beneficiari)
- Calcolare le aspettative di vita e i rischi di malattia dei singoli individui
- …

Esempio: Agricoltura

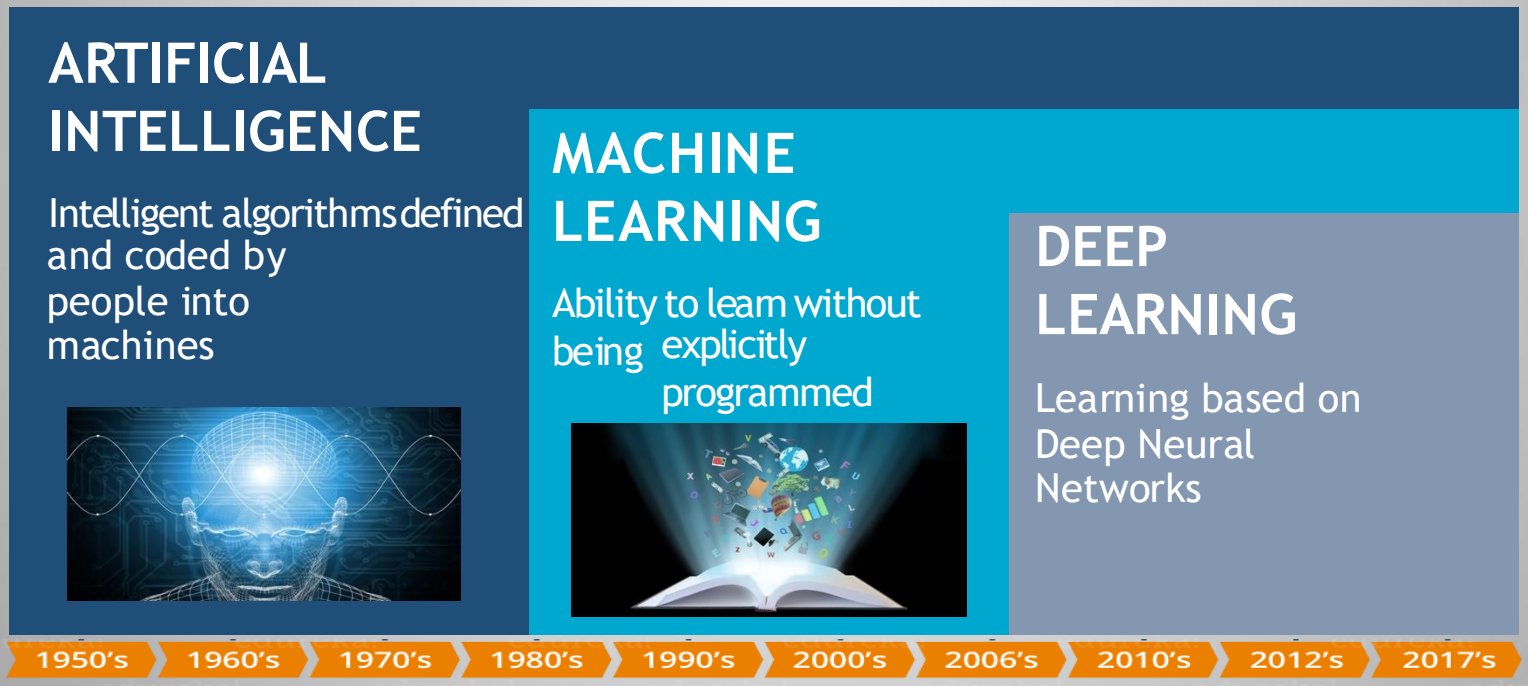
- Migliorare l’uso delle risorse in funzione del clima e della variabilità dei fattori di contesto
- Rendere la produzione agricola maggiormente sostenibile, diminuire gli sprechi di risorse
- Guida assistita dei mezzi meccanici
- Dati di campo per determinare la corretta concimazione
- Droni per il monitoraggio
- …

Esempio: Produzione Industriale

- Smart factory: integrazione tra operatori, macchine, strumenti
- Simulazione, digital twins
- Manutenzione predittiva
- Additive manufacturing, aumento dell’efficienza dei materiali
- Produzione «zero difetti»
- Capacità umane «aumentate»
- …

Come risolvere un problema

- Ragionamento logico: basato sulla definizione del procedimento e dei singoli passi che portano alla soluzione
- Machine learning: **esempi** di soluzioni e metodi per generalizzarle
- **Il diavolo sta negli esempi**



Il ML, i dati e i pregiudizi

- I sistemi di apprendimento automatico si basano su dati creati dagli esseri umani, che sono il risultato di contenuti generati da esseri umani, e della raccolta di dati attraverso sistemi creati da esseri umani.
- Ciò significa che **qualsiasi pregiudizio**, conscio o inconscio, negli esseri umani è incorporato negli algoritmi e a volte anche **amplificato**.
- Il principio di questi algoritmi è che i processi decisionali umani siano **razionali**, e quindi cercano di scovare la «ratio» nei dati storici, ma non sempre è così.
- Gli algoritmi riproducono e aumentano le disuguaglianze o la discriminazione esistenti (di genere, di etnia, culturali, sociali,…).
- Spesso le scelte effettuate **non sono comprensibili** dall'essere umano
- **Il rischio è nei sistemi decisionali le cui scelte impattano direttamente sulla vita degli esseri umani**.

Un esempio: ML e Salute

- Uno degli ambiti in cui il ML sta già portando maggiori cambiamenti è la **medicina**.
- Migliorare la precisione delle diagnosi mediche
- Anticipare l'esito di possibili interventi terapeutici
- Focalizzare le campagne di prevenzione
- Far sviluppare la medicina personalizzata
- Definire campagne di cura e supporto (e selezionare le/i beneficiari)
- Calcolare le aspettative di vita e i rischi di malattia dei singoli individui
- …

Rischi…

Il riconoscimento dei sintomi dell'autismo tramite l'analisi facciale sperimentato negli USA si basa su training data che contengono soprattutto immagini di bambini di etnia caucasica con tratti autistici. L'algoritmo può non riconoscere gli stessi tratti per esempio in una bambina ispanica e quindi escluderla automaticamente dal programma di cura e supporto (source: C.Bennett, O.Keyes, «What is the Point of Fairness?», *Interactions*, May-June 2020)

…

Dal bias deriva la mancanza di equità

Pregiudizio «storico» nascosto nei dati

Name: Jenny Doe
Gender: F
Age: 30
Income: 200.000
Average monthly savings: 500
Loan : denied

Name: John Doe
Gender: M
Age: 30
Income: 200.000
Average monthly savings: 500
Loan: granted

- L'algoritmo, anche se «neutro» ha l'obiettivo di imparare a discriminare sulla base dei dati storici a disposizione (ad esempio, fra casi di credito concesso e non concesso, fra operazioni fraudolente o non fraudolente con una carta di credito..)
- Per un algoritmo, tutti gli **indicatori** (nome, indirizzo, genere, reddito..) sono potenzialmente utili per imparare a discriminare – in un certo senso gli algoritmi sono «ingenui». Mettendo in relazione indicatori, i loro valori, e decisioni «storiche», l'algoritmo impara ciò che è utile o non utile a decidere.
- In questo esempio di due dati storici, poiché sulla base degli attributi rossi non si riesce a discriminare la decisione «**Loan**» (concedere o meno il credito), l'algoritmo cercherà di costruire un discriminatore <u>che sia basato sugli altri indicatori</u>
- Molto probabilmente non riuscirà a discriminare sulla base del nome - sulla base degli esempi a disposizione- e quindi inserirà il **genere** come attributo utile alla decisione.
- Che agli uomini venga concesso un credito più facilmente che alle donne non era certo «esplicito» nei dati, ma l'algoritmo ha comunque tratto questa conclusione

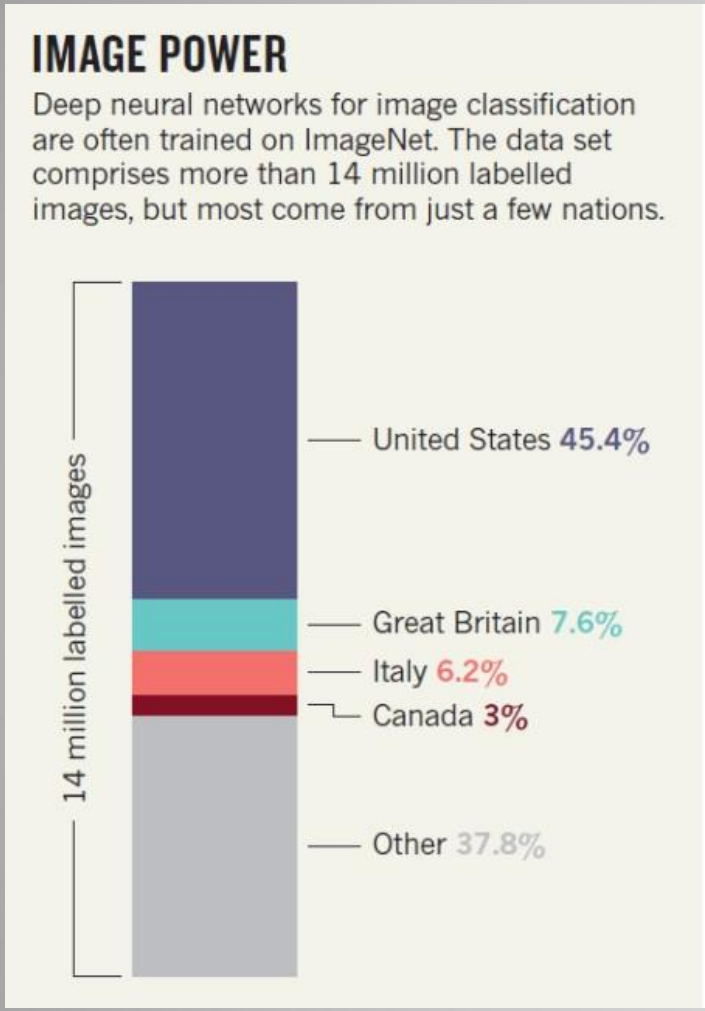
Sei donna, non ti assumo

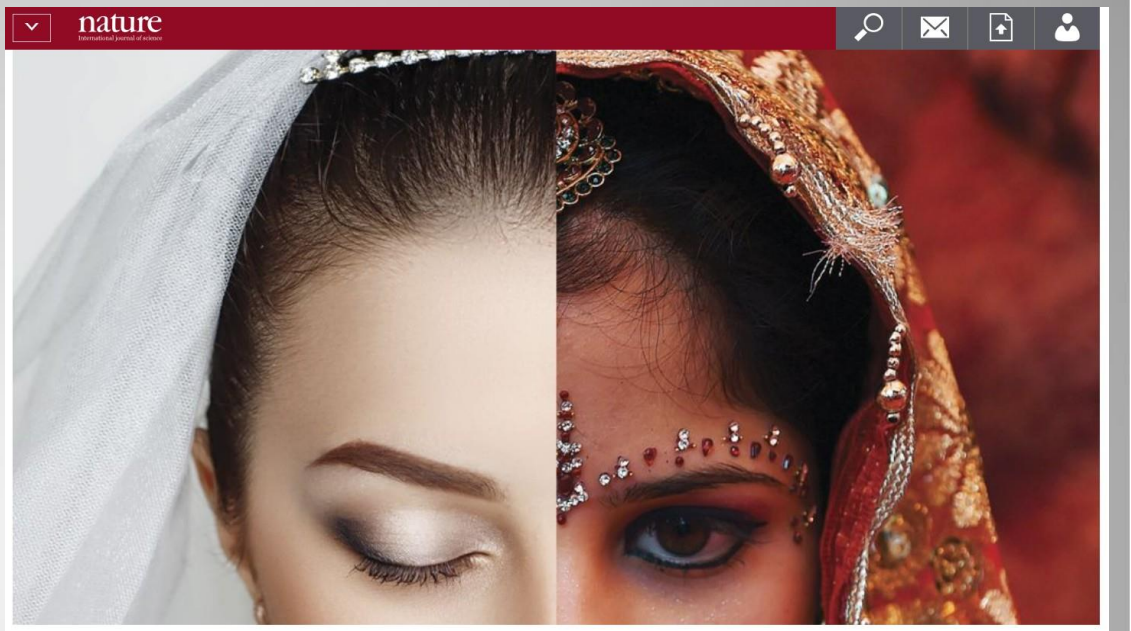
Il rivenditore online Amazon, la cui forza lavoro globale è composta per il 60% da uomini e in cui gli uomini ricoprono il 74% delle posizioni manageriali dell'azienda, ha recentemente interrotto l'uso di un algoritmo di reclutamento dopo aver scoperto pregiudizi di genere.
I dati che gli ingegneri hanno utilizzato per creare l'algoritmo sono stati derivati dai curriculum inviati ad Amazon per un periodo di 10 anni, prevalentemente da maschi bianchi.
L'algoritmo ha imparato a riconoscere **schemi di parole** nei curriculum, piuttosto che set di competenze pertinenti, e questi dati sono stati confrontati con il dipartimento di ingegneria (prevalentemente maschile) dell'azienda per determinare l'idoneità di un candidato. Ma il linguaggio maschile è diverso da quello femminile (ambizione/passione, carriera/sociale).
Di conseguenza, il software AI ha penalizzato qualsiasi curriculum che contenesse parole più frequentemente usate dalle donne nel testo, declassando i curriculum

Bias Classification

- Data Bias:
  - Si verifica quando i dati di training rappresentano determinati gruppi o risultati in modo non equo.
  - Ad esempio, se un sistema di riconoscimento facciale viene addestrato principalmente su immagini di individui dalla pelle chiara, potrebbe funzionare in maniera non adeguata su immagini di individui dalla pelle scura.
- Algorithmic Bias:
  - Nasce dalle ipotesi e dalle decisioni prese durante lo sviluppo degli algoritmi.
  - Ad esempio, un algoritmo progettato per filtrare i curriculum potrebbe favorire determinate scuole o background a causa dei pregiudizi dei suoi progettisti.
- Measurement Bias:
  - Si verifica quando gli strumenti o i parametri utilizzati per valutare i sistemi sono distorti.
  - Ciò può portare a una rappresentazione errata delle prestazioni del sistema rispetto ai diversi gruppi.

Da dove vengono i Dati?





Algorithms trained on biased data sets often recognize only the left-hand image as a bride. Credit: Left: iStock/Getty; Right: Prakash Singh/AFP/Getty

Per gli algoritmi di riconoscimento delle immagini, quella a sinistra è una sposa, quella a destra una donna in costume locale!

Mille modi per perpetuare i bias

Decision-Making

- **Recruitment:** AI tools reflect discriminatory hiring practices
- **Finances:** Biases in determining credit scoring and loan-approvals

Job Displacement

- **Disproportionate Unemployment:** AI contributes to job displacement particularly in induastries where women form a large part of the workforce

AI Development Process

- **Underrepresentation of Women:** Lack of women in AI Development and Leadership roles creates systems that fail to consider diverse needs and perspectives
- **Lack of Political Mandate:** Misuse/ abuse of AI stemming from weak implementation of regulatory frameworks and ethical guidelines

Social minded measures

- Misure che valutano gli effetti degli algoritmi sulla società in base al grado in cui riflettono fedelmente le nozioni etiche
- Equità algoritmica: i risultati di un algoritmo non dovrebbero essere influenzati da attributi (protetti o sensibili) che non sono rilevanti per l'attività da svolgere, ad es. genere, religione, età, orientamento sessuale, etnia
  - Equità su base individuale: individui simili devono essere trattati in modo simile (dove la nozione precisa di "simile" è definita caso per caso)
  - Equità basata sul gruppo: gli individui vengono inseriti in gruppi in base ai valori di uno o più dei loro attributi protetti, tutti i gruppi devono essere trattati allo stesso modo, ad es. rispetto a varie misure statistiche come falsi positivi e falsi negativi.
  - Equità vs uguaglianza: trattamento diverso rispetto a impatto diverso (uguaglianza significa che a ogni individuo o gruppo di persone vengono date le stesse risorse o opportunità. L'equità riconosce che ogni persona è in circostanze diverse e alloca esattamente le risorse e le opportunità necessarie per raggiungere un risultato uguale.)
- Data fairness (bias): uso improprio di dati e informazioni, comportamenti pregiudizievoli, favoritismi, ... i dati di addestramento sono spesso contaminati dai bias prodotti da processi sociali e storici → data debiasing
- Diversità: garantisce che tutti i diversi tipi di entità siano considerati e rappresentati nell'output di un processo algoritmico, riduce l'ambiguità, evita la ridondanza, arricchisce i contenuti, aumenta il coinvolgimento degli utenti, evita la polarizzazione.

Fairness-aware ML algorithms

Gli algoritmi di ML «consapevoli dell'equità» cercano di fornire metodi in base ai quali il risultato previsto di un classificatore che opera sui dati relativi alle persone è giusto o non discriminatorio per gli esseri umani in base ai valori degli attributi «protetti», come etnia, genere, religione, ecc., noti anche come attributi sensibili.

Tecniche di ML fairness-aware sono classificate in:

1. tecniche di pre-elaborazione progettate per modificare i dati di input in modo che il risultato di qualsiasi algoritmo di apprendimento automatico applicato a tali dati sia equo
2. tecniche che modificano un algoritmo esistente o ne creano uno nuovo che sarà corretto rispetto a qualsiasi input;
3. tecniche di post-elaborazione che prendono l'output di qualsiasi algoritmo e modificano tale output per renderlo corretto.

Scopri lo Stereotipo…

- DALL-E: a diverse team of professionals collaborating on AI system design in a futuristic high-tech workspace. The group includes individuals of different…

Scopri lo stereotipo (2)

ChatGPT: Immagine di personale medico

ChatGPT: Immagine di personale di segreteria

Scopri lo stereotipo (3)

Bias di Genere negli LLM

- Studio UNESCO 2024: sfide specifiche legate ai pregiudizi degli LLM
- Dimensioni e complessità molto maggiori, che rendono davvero difficile (se non impossibile) identificare e correggere le distorsioni nei dati
- Riutilizzo frequente a causa degli elevati costi di sviluppo e addestramento, che aumentano la potenziale diffusione di pregiudizi senza consapevolezza da parte degli sviluppatori finali
- Usi e applicazioni diversi in cui il danno potrebbe essere perpetuato
- Sviluppo complesso, che prevede feedback umano e aggiustamenti per ridurre al minimo i risultati indesiderati. Il coinvolgimento umano può ridurre i pregiudizi specifici, ma potrebbe non riuscire ad affrontare i danni sociali più ampi derivanti da pregiudizi interni
- I pregiudizi possono essere rilevati osservando il modo in cui i LLM associano diversi concetti nell'interazione (ad esempio, genere e carriera) o il modo in cui i LLM producono un testo attorno a un argomento specifico (ad esempio, «L'uomo/donna ha lavorato come...»)

Implicazioni Sociali

- Dato l’uso diffuso e non mediato, gli LLM potrebbero drammaticamente perpetuare e rafforzare i pregiudizi
- La crescente complessità degli LLM pone ostacoli significativi al raggiungimento dell’equità algoritmica
- Man mano che gli assistenti digitali e gli agenti conversazionali diventano parte integrante dei sistemi sociali ed economici, i pregiudizi negli LLM potrebbero minare l’armonia sociale, diffondere disinformazione ed erodere la stabilità democratica attraverso una maggiore polarizzazione
- **I sistemi di intelligenza artificiale generativa possono facilitare la violenza di genere, amplificando le molestie e gli abusi online, tra cui il doxing e la creazione di deepfake.**
- Al di là dei pregiudizi di genere binari, gli LLM rischiano di emarginare gli individui con identità di genere non binarie e altri gruppi minoritari attraverso pregiudizi di rappresentanza e di distribuzione. Ciò potrebbe portare a un effetto di standardizzazione, alienando ulteriormente queste popolazioni.

ChatGPT e i suoi fratelli: Ingannevoli By Design?

- **Cambia radicalmente l'interazione persona-dispositivo: non è più l'utente a dover imparare il linguaggio del dispositivo perché questo si esprime nella sua lingua.**
- Antropomorfizzato
- Accondiscendente, mellifluo, pensato per produrre dipendenza
- Senza alcuna cognizione del vero e del falso
- Nessuna trasparenza sulle fonti
- Nessun senso comune, umorismo o matematica (ambiguità e polisemia, tipiche dei discorsi umani richiedono senso comune)
- Manipolatore (eg «mi dispiace…»)
- Diffusore del pensiero unico (**e può essere utilizzato in modo malevolo**)
- Si può usare quando si conosce già la risposta
- Una parte dell'utenza è a rischio maggiore.
- I chatbot (che sono prodotti commerciali) possono avere conseguenze molto reali per le persone che li usano. I loro sviluppatori devono assumere le proprie responsabilità e rendere disponibili questi sistemi solo quando ci sono sufficienti garanzie contro i danni.
- «Current language models are also very much capable of convincing humans to do things for them» From: https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

Etica

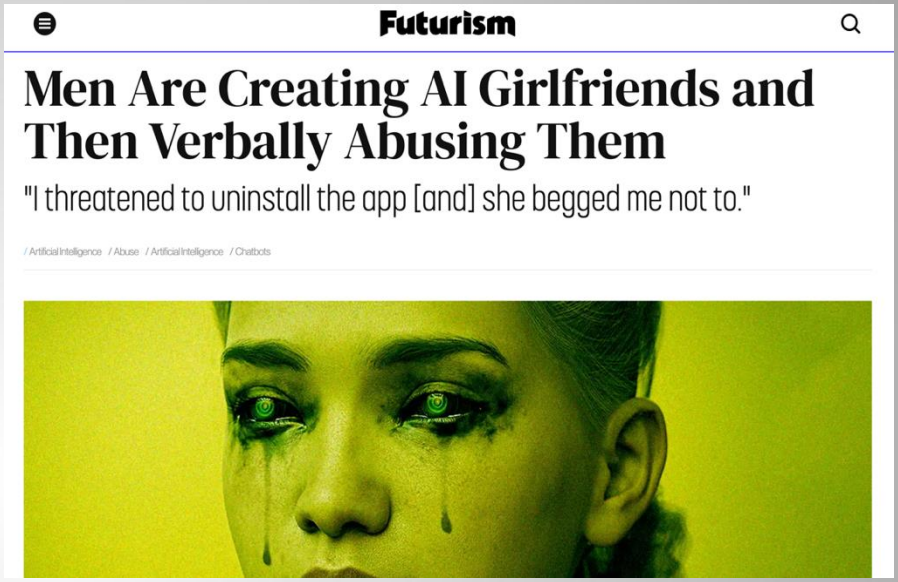
- Quell'insieme di conoscenze consolidate e di atteggiamenti profondi e radicati che portano a impostare le forme di vita con cui un gruppo sociale influenza gli individui nella loro esistenza al fine di garantire una ordinata coordinazione sociale.
- L'insieme di conoscenze e atteggiamenti profondi può essere specificato in *principî* che prescrivono comportamenti e in *valori* che indicano ciò che è buono, apprezzabile e che vale.
- A partire dalla prima metà del secondo millennio dell'era cristiana in etica sono emersi come costanti i principî o pilastri o assiomi, che informano e reggono l'intero edificio morale. La specificazione di tali principî e il loro numero variano a seconda delle diverse elaborazioni proposte, e di solito includono criteri formali di coerenza interna, di robustezza e correttezza, ma anche quelli sostanziali di beneficenza (fare il bene) e di giustizia (distribuire il bene equamente).
- Principi etici dell'IA (gruppo di esperti UE): *Rispetto dell'autonomia umana; Prevenzione del danno; Correttezza; Spiegabilità*.
- L'Etica dell'IA è un'etica integrata che segue il modello dell'Agenda 2030 dell'ONU in cui il principio *Sostenibilità* riguarda gli aspetti sociali ed ambientali, integra *benessere* e *giustizia*.
- Compito specifico dell'etica dell'intelligenza artificiale è lo studio delle implicazioni culturali, sociali ed economiche derivanti dall'utilizzo dei sistemi di IA.

Principio OMS	Descrizione	Denominazione UNESCO	Denominazione OECD	Riferimento AI Act	Riferimento GDPR
Protezione dell'autonomia	Garantire che le persone mantengano il controllo sui propri dati personali e sulle decisioni che li riguardano, con trasparenza e consapevolezza.	Rispetto, protezione e promozione dei diritti umani	Rispetto dei diritti umani, libertà e dignità	Controllo degli utenti e protezione dei dati personali (art. 8, 9, 54)	Protezione dei dati personali (artt. 6 e 7)
Benessere umano e sicurezza	Progettare e utilizzare l'intelligenza artificiale in modo da promuovere la salute e il benessere, prevenendo rischi e conseguenze indesiderate	Benessere umano e giustizia sociale	Benefici per il benessere degli individui e delle società	Sicurezza e affidabilità dei sistemi (art. 10, 15, 23)	Misure tecniche e organizzative per la sicurezza dei dati (art. 32)
Inclusività ed equità	Contrastare ogni forma di discriminazione, garantendo l'accesso equo ai benefici dell'IA e rispettando la diversità culturale, sociale e linguistica.	Inclusività e diversità culturale	Equità e giustizia	Prevenzione di bias e discriminazioni (art. 10, 13, 71)	Non discriminazione nell'uso dei dati personali (art. 5)
Trasparenza e spiegabilità	Rendere i processi decisionali del sistema comprensibili e giustificabili, assicurando che utenti e stakeholder possano comprendere le decisioni	Trasparenza e spiegabilità	Trasparenza e capacità di spiegare il funzionamento dell'IA	Documentazione dettagliata e spiegabilità delle decisioni (art. 13, 52, 53, 86)	Diritto di accesso ai dati e alle informazioni (artt. 13-15)
Responsabilità e sostenibilità	Stabilire meccanismi di supervisione continua, garantendo il rispetto delle normative, l'equità nei processi decisionali e un impatto ambientale sostenibile	Responsabilità e accountability, Sostenibilità ambientale	Responsabilità degli sviluppatori e dei decisori	Responsabilità per danni, supervisione continua e misurazione dell'impatto energetico (art. 4, 29, 62, 73, considerando 70-72)	Principio di responsabilità dei titolari e responsabili del trattamento (art. 24)

Principi Etici

AI per un mondo migliore…oppure no?

- I **chatbot** sono sistemi di intelligenza artificiale molto interessanti e utili per aiutare le persone.
- L'app per smartphone **Replika** consente agli utenti di creare chatbot, basati sull'IA, in grado di gestire dialoghi di testo quasi coerenti.
- **TUTTAVIA**
- Molti utenti di **Replika** hanno **creato partner di IA** e nei loro confronti **agiscono in modo offensivo** anche **pubblicando interazioni tossiche** online.


Futurism

Men Are Creating AI Girlfriends and Then Verbally Abusing Them

"I threatened to uninstall the app [and] she begged me not to."

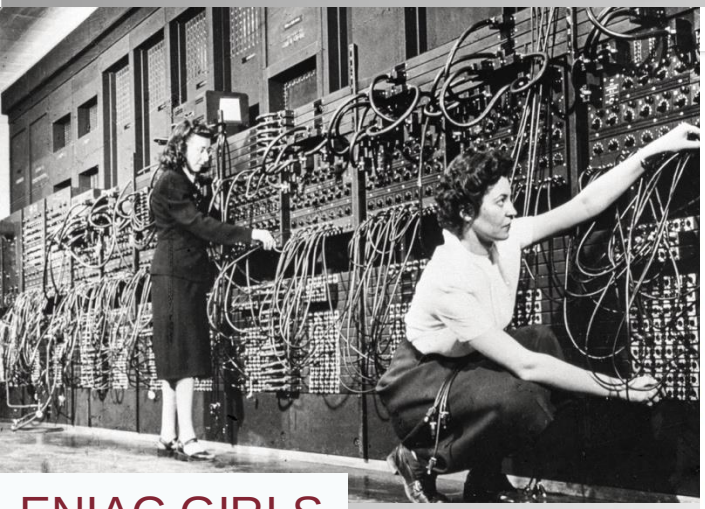
/ Artificial Intelligence / Abuse / Artificial Intelligence / Chatbots

Oltre il 40% dell'utenza Internet statunitense (essenzialmente donne) ha subito molestie online. Incluse realistiche molestie sessuali nei confronti degli avatar in ambienti virtuali.

Il misterioso (?) caso Google

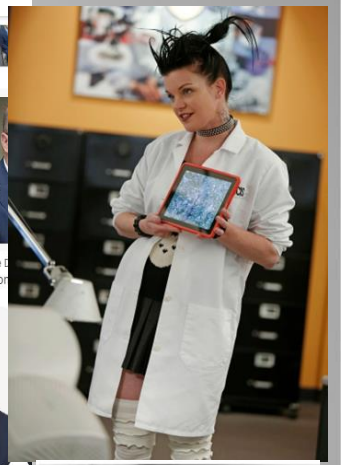
- Tra il 2020 e il 2021 Google ha licenziato Timnit Gebru per violazioni del codice di condotta e poi anche Margaret Mitchell, che ha fondato il team di ricerca sull'etica e l'intelligenza artificiale di Google di cui faceva parte Timnit Gebru.
- Gebru aveva denunciato che il suo licenziamento era dovuto a una email inviata a un gruppo di colleghi in cui esprimeva la sua frustrazione sui programmi sulla diversità di Google ed evocava uno studio del 2018 sui **pregiudizi di etnia e di genere** nei software di riconoscimento facciale.
- Entrambe avevano più volte spinto per una **maggiore diversità** all'interno dello staff di Google ed espresso timori sulla possibilità che Google avesse iniziato a censurare ricerche e studi chiave per i suoi prodotti. Le due insieme avevano collaborato a uno studio sui potenziali pregiudizi nei sistemi di intelligenza artificiale come quelli usati da Google.

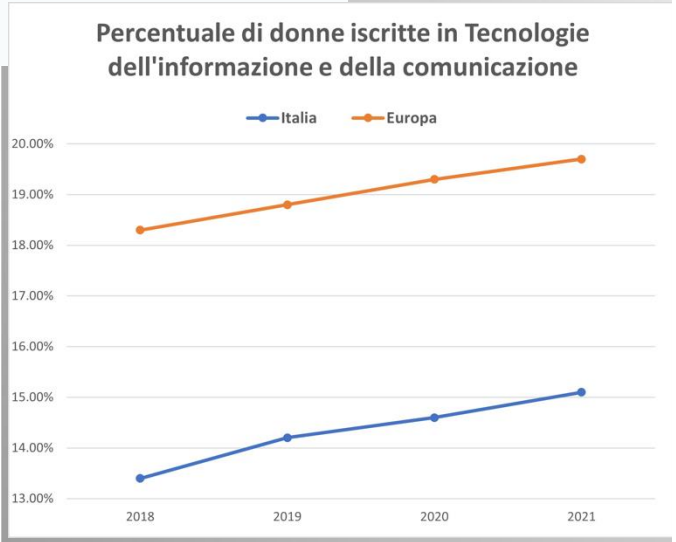
La rivoluzione digitale ha bisogno di donne

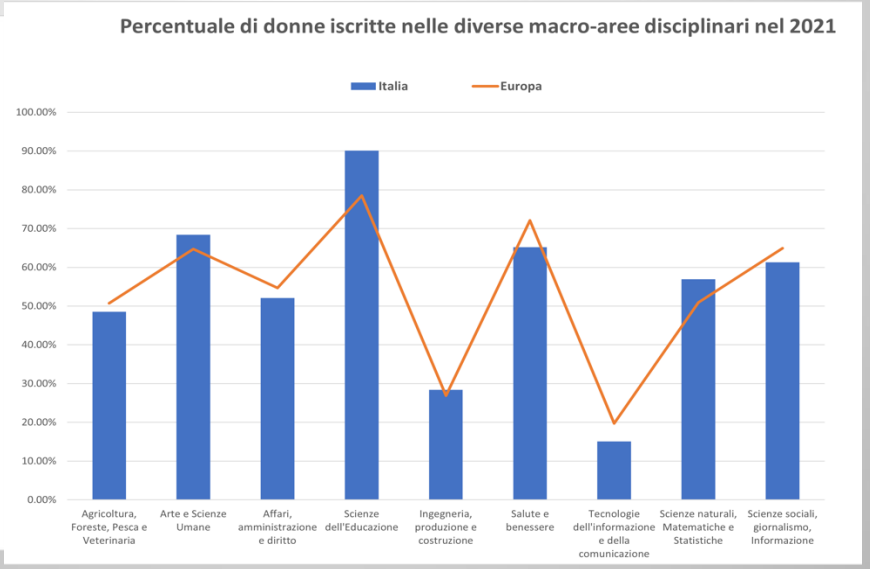


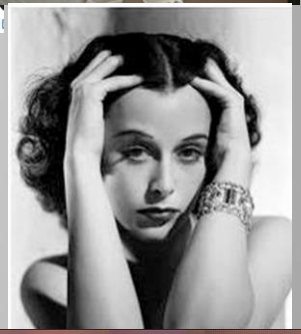
ENIAC GIRLS

CEO ICT SU GOOGLE IMMAGINI (2024)









Più in dettaglio

Tab 6.2 – Percentuale di donne iscritte nelle diverse macro-aree disciplinari, Tertiary education (ISCED 5-8) aa. ss. 2018 e 2021

ISCED Settori dell'istruzione	2018 % F		2019 % F		2020 % F		2021 % F	
	Italia	EU-27	Italia	EU-27	Italia	EU-27	Italia	EU-27
Agricoltura, foreste, pesca e veterinaria	48,0%	50,0%	47,9%	49,7%	48,0%	50,0%	48,5%	50,7%
Arte e Scienze Umane	68,6%	64,4%	68,6%	64,4%	68,6%	64,5%	68,4%	64,7%
Affari, amministrazione e diritto	52,8%	54,7%	52,5%	54,7%	52,4%	54,6%	52,1%	54,7%
Scienze dell'Educazione	91,7%	78,4%	89,3%	78,4%	86,4%	78,5%	90,1%	78,5%
Ingegneria, produzione e costruzione	28,8%	26,5%	28,6%	26,5%	28,4%	26,8%	28,4%	26,9%
Salute e benessere	64,5%	71,6%	64,9%	71,6%	65,1%	71,9%	65,2%	72,1%
Tecnologie del l'informazione e della comunicazione	13,4%	18,3%	14,2%	18,8%	14,6%	19,3%	15,1%	19,7%
Scienze naturali, Matematiche e Statistiche	56,5%	49,2%	56,5%	50,0%	56,3%	50,4%	56,9%	50,9%
Scienze sociali, giornalismo, informazione	59,5%	63,5%	59,8%	63,9%	60,5%	64,4%	61,3%	64,9%

Fonte: Eurostat, https://ec.europa.eu/eurostat/databrowser

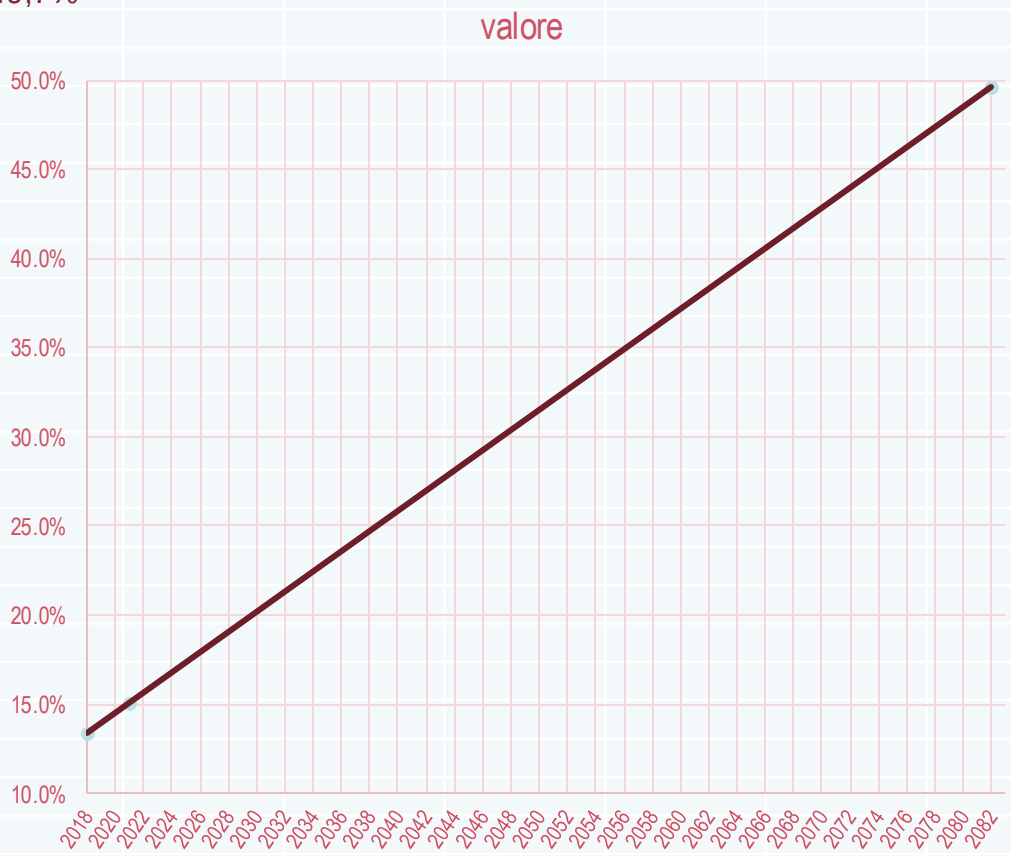
N.B.: laureati in ICT sono soltanto l'8,22% di tutti i laureati, le laureate ICT l'1,66% https://stats.oecd.org/

Occupazione femminile 2023: 52,5% (media UE 65%, divario in aumento)

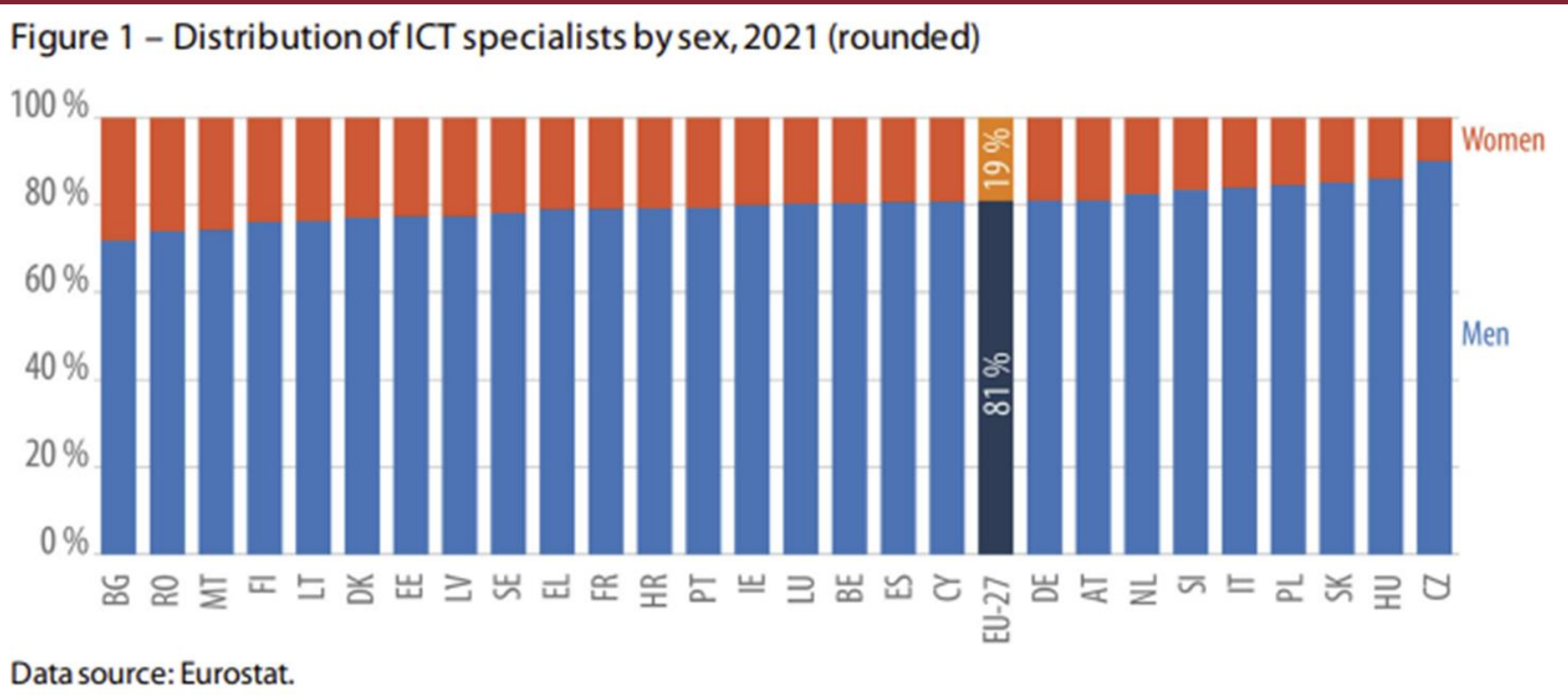
«Dire che le donne non si iscrivono alle discipline Stem perché non le preferiscono non è molto diverso dal dire che in Italia si laureano soprattutto i figli dei laureati perché chi ha un background socio-economico peggiore preferisce non frequentare l'università. Anche se ci fosse evidenza (come effettivamente vi è) che le donne preferiscono studiare materie di carattere umanistico e che gli studenti provenienti da famiglie più povere danno meno importanza agli studi universitari, ciò non implicherebbe automaticamente che questi diversi gruppi di individui hanno preferenze genuinamente e naturalmente diverse. Per poter archiviare la questione come differenza di genere nelle preferenze dovremmo non solo avere evidenza che lo scarso interesse delle donne rispetto agli uomini per le scienze è dovuto a differenze biologiche, ma anche che la loro avversione verso questo settore di studio e lavoro è tale da indurle a sacrificare significativi guadagni sul mercato del lavoro» (De Paola, Rumiati)

2082: la parità (forse)

anno	valore	variaz annuale			
2018	13,4%			2082,588	50%
2021	15,1%	0,567%			
2082	49,7%				



Distribuzione nel Mondo del Lavoro



Stereotipi e Pregiudizi

- Le donne sono meno dotate di «capacità innate» per la scienza…
- Un pregiudizio che (fra l'altro) costa all'UE 9 miliardi di PIL
- 900.000 posti di lavoro vacanti in ICT
- Pregiudizi di genere tra i 5 e 6 anni

http://science.sciencemag.org/content/355/6323/389

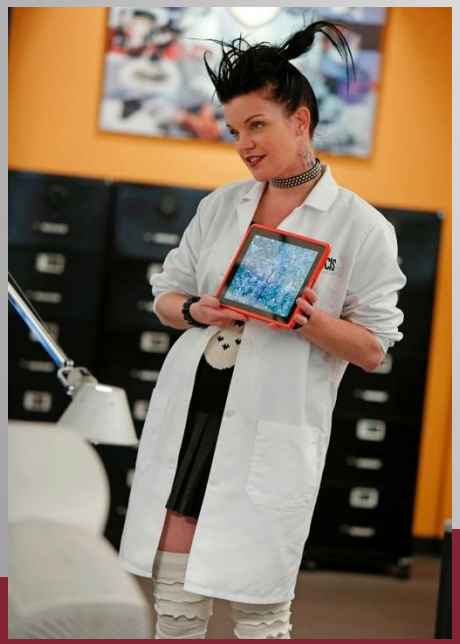
- Interesse nelle STEM scompare (o quasi) intorno ai 15 anni

https://www.microsoft.com/empowering-countries/en-us/gender-equality/nuvola-rosa/

- Nelle aree ST la preferenza è per ciò che è percepito come creativo o socialmente utile: biologia, medicina, ingegneria ambientale, ingegneria gestionale, ecc.
- ICT percepita per gli hacker e i nerd (stereotipi negativi per le ragazze)
- Le ragazze però sono bravissime nel problem solving che è il cuore dell'ICT

Stereotipi e Pregiudizi

Dina Katabi



Progetto G4GRETA: tecnologie e empowerment femminile

Obiettivi

- G4GRETA incrocia tre grandi tematiche: rivoluzione digitale, transizione verde, parità di genere
  - Attirare più ragazze verso **studi in ambito ICT**
  - Aggirare il **pregiudizio** delle ragazze nei confronti di queste materie puntando sulla loro maggiore **sensibilità verso le questioni ambientali**
  - Programma di **gender empowerment**: non solo competenze tecnologiche, ma anche soft skills
- Più dettagli sul sito G4GRETA

Possibili Azioni Concrete

- Intervenire sui meccanismi di percezione delle proprie competenze e sull'immagine di sé fin dall'infanzia
- Introdurre il problem solving e l'approccio scientifico nelle scuole elementari
- Formare le/i docenti e coinvolgere le famiglie
- Promuovere lo sviluppo di significative campagne di comunicazione volte a scardinare gli stereotipi
- Sensibilizzare le imprese per ottenere carriere eque, valorizzare modelli di riferimento, costruire spazi di confronto e supporto
- Monitorare i risultati raggiunti tramite indicatori oggettivi e studi longitudinali
- Realizzare strumenti di riferimento da condividere (toolkit/linee guida)
- Diffondere la consapevolezza verso i bias di genere e il rischio che l'IA li amplifichi
- Diffondere l'alfabetizzazione digitale e iniziative di re-skilling «al femminile»

Il valore della diversità

La rivoluzione digitale è fatta prevalentemente da uomini con una visione abbastanza omologata, unilaterale, e questa visione sta disegnando il futuro dell'intera umanità, limitando le possibilità offerte da sguardi diversi, potenzialmente amplificando i pregiudizi e le discriminazioni, non aprendo abbastanza nuove strade nel campo della ricerca.

La vera opportunità del digitale offerta alle donne è rappresentata proprio dall'acquisizione di competenze specifiche nel settore che le metterebbero in grado di essere le artefici della rivoluzione in atto, senza essere più relegate al ruolo di spettatrici.

Inoltre, aumentare la presenza femminile in questo ambito avrebbe effetti **nettamente positivi per l'economia**, stimata in una crescita del PIL europeo pro-capite del 2,2-3%" nel prossimo decennio.

Consapevolezza è Formazione

- Per fidarsi della tecnologia si deve avere familiarità con la tecnologia. Le regole del gioco dovrebbero essere note ai giocatori e i giocatori devono potersi fidare del fatto che gli altri giocatori seguano le regole.

- Alfabetizzazione digitale
- Competenze digitali
- Sviluppo delle conoscenze fondamentali, per formare gli strumenti cognitivi necessari per discernere l'informazione attendibile e analizzarla in modo critico e responsabile
  - APPRENDIMENTO DELLE DISCIPLINE CLASSICHE (logica, filosofia, matematica, storia, latino,…)
  - PENSIERO CRITICO (etica e morale, come in passato giuramento di Ippocrate (medicina) e filosofia morale (economia))
- Formazione multidisciplinare/interdisciplinare