Decisioni senza decisori? Nuovi scenari tecnologici e questioni aperte tra etica e psicologia

13° Workshop CIPA - Innovazione IT e banche

Nuove tecnologie nei processi aziendali e di business delle banche:

Intelligenza Artificiale generativa e Distributed Ledger Technology

Francesco La Barbera
Università degli Studi di Napoli Federico II

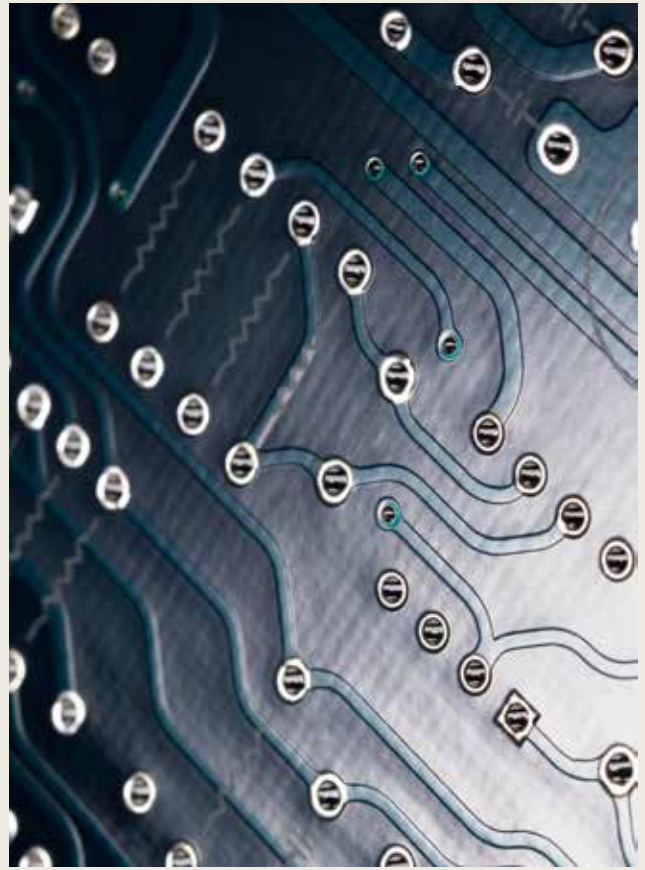
L’intelligenza artificiale

L'IA comprende algoritmi, metodologie e tecnologie che consentono alle macchine di simulare funzioni cognitive simili a quelle umane.

Alla base c’è il machine learning (ML), che consente ai computer di imparare e fare previsioni (o prendere decisioni) basate sui dati, migliorando nel tempo (DL).

Queste tecnologie rendono possibile l'IA generativa, ovvero modelli di IA progettati per **generare** dati nuovi, spesso indistinguibili dai contenuti creati dagli umani.

L'IA oggi non si limita all'esecuzione di compiti, ma consente alle macchine di **apprendere** e **adattarsi**. La proliferazione dei big data, l'aumento della potenza di calcolo e i progressi degli algoritmi hanno alimentato questa evoluzione dell’IA.



IA: una classificazione rispetto alle funzioni cognitive

Cognitive Function	Purpose	Level
Meta-cognitive and self-aware	Modelling mental states of agents, including own mental states, based on "self" concept	Highest
Reflective	Modelling internally the environment and behaviour of entities and objects in it	High
Proactive or Deliberative	Reasoning, planning, exploration, and decision making	Middle
Reactive or Adaptive	Sub-cognitive forms of learning and adaptation	Low
Reflexive	Pre-programmed behavioural responses	Lowest

Competenza + Assenza di interesse personale =

Mito della neutralità

Decisioni senza decisori?

Problemi di funzionamento

Problemi di sicurezza e manipolazione

Problemi di distorsioni sistematiche della realtà (Bias)

Problemi di responsabilità

Problemi di fiducia e trasparenza

DeepL Translate: Il miglior tra | screenshot mac - Cerca con | Usa, primo incidente mortale

repubblica.it/motori/sezioni/sicurezza/2016/07/01/foto/usa_primo_incidente_mortale_di_auto_a_guida_autonoma-143175770/1/

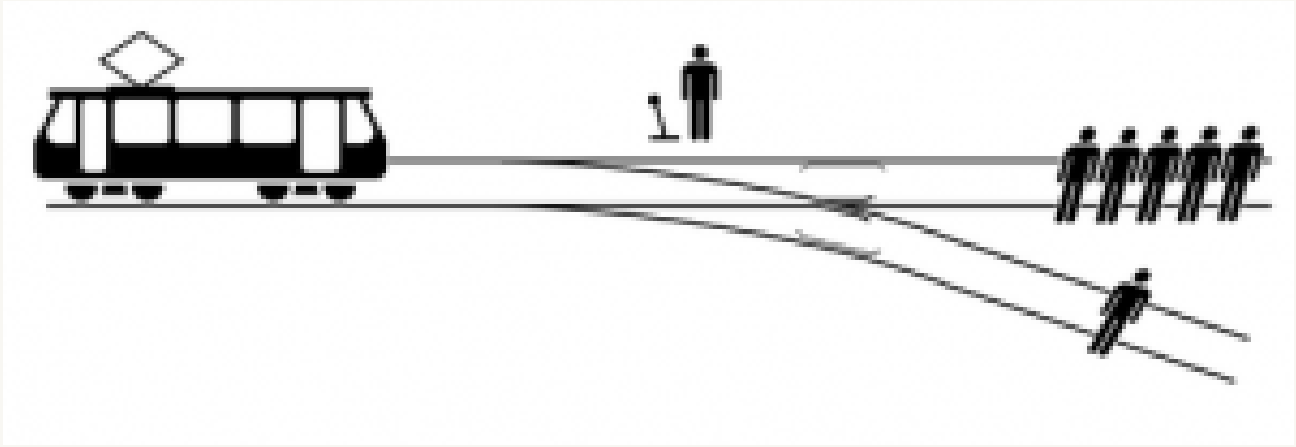
Riavvia per aggiornare

Problemi di funzionamento

Usa, primo incidente mortale di auto a guida autonoma

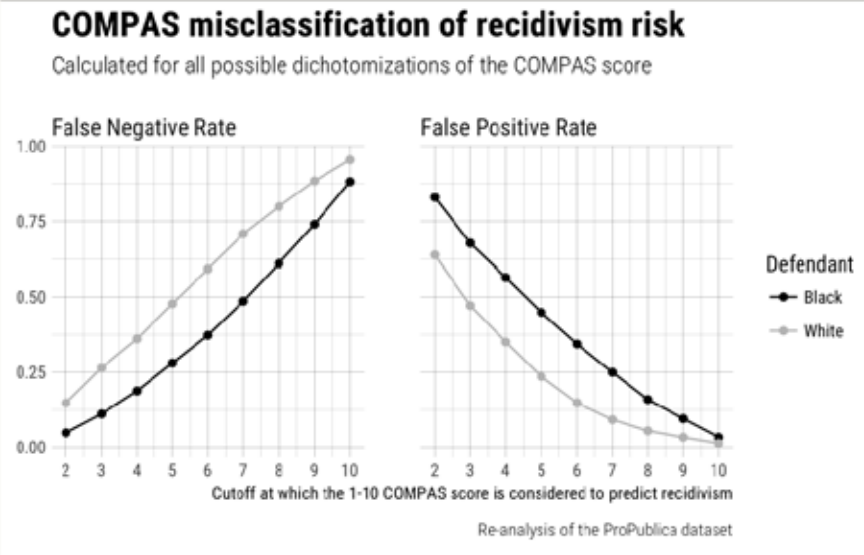
01 Luglio 2016

Si è verificato in Florida il 7 maggio scorso il primo incidente stradale mortale di un'auto a guida autonoma negli Stati Uniti. La vittima era il conducente di una Tesla S, che aveva inserito il sistema di guida



Problemi di responsabilità

Problemi di distorsioni sistematiche della realtà (Bias)



DISCRIMINAZIONI DIGITALI

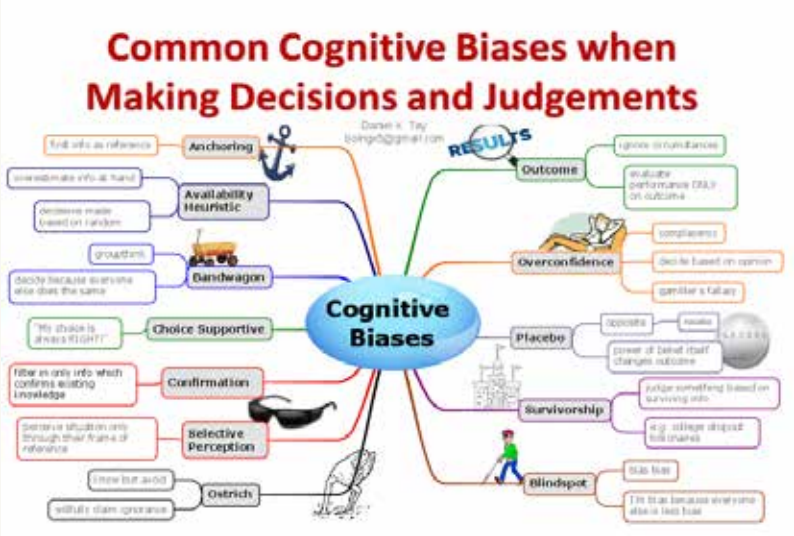
Amazon e l'intelligenza artificiale sessista: non assumeva donne

Secondo quanto riporta Reuters, i software utilizzati per individuare i migliori talenti in campo tecnologico prediligevano curriculum maschili, basandosi sull'andamento delle assunzioni nei 10 anni precedenti. Il progetto è stato chiuso nel 2017

di Andrea Federica De Cesco

Human heuristics/AI heuristic algorithms

»Although optimal solutions can be obtained using exact methods, the computational time is prohibitive. Since heuristic methods often produce near optimal solutions in a reasonable amount of computational time… (Desrochers et al., 1992)



Problemi di sicurezza

Data poisoning

Violazione

Deep fake e manipolazione

L'intelligenza artificiale è immune da attacchi hacker?

Falso: E' esposta come tutte le tecnologie, è possibile anche comprometterne l'addestramento

Redazione ANSA

24 maggio 2023 - 13:42

CONDIVIDI

L'IA diventa hacker: individuato uno script PowerShell creato con strumenti generativi

L'uso dell'IA come strumento a corredo delle attività di hacker e criminali informatici cambia il panorama della sicurezza informatica, imponendo un nuovo approccio all'analisi e alla ricerca delle minacce

di Andrea Bai pubblicato il 11 Aprile 2024, alle 10:31 nel canale SICUREZZA

Proofpoint

Un attore di minaccia noto come TA547 ha condotto una serie di attacchi a diverse aziende in Germania utilizzando uno **script PowerShell** che è presumibilmente stato **creato con l'aiuto di un sistema di intelligenza artificiale generativa come ChatGPT, Gemini o CoPilot**.

L'aggressore ha orchestrato una campagna di phishing, rilevata durante il mese di marzo, che ha preso di mira decine di organizzazioni tedesche allo scopo di diffondere il malware **Rhadamanthys**, un cosiddetto "information stealer".

E' stata la società di sicurezza informatica Proofpoint ad aver attribuito attribuito l'attacco a TA547, un broker di accesso iniziale (IAB) attivo dal 2017 e noto per fornire una varietà di malware per sistemi Windows e Android. TA547 è anche noto con il nome di Scully Spider.

Data poisoning nel settore bancario

split view poisoning

manipolazione di URL scaduti su Internet per introdurre dati finanziari falsificati nei sistemi di IA, con conseguenti imprecisioni nel rilevamento delle frodi e nella valutazione del credito.

frontrunning poisoning

manomissione di archivi di dati affidabili per inserire falsi indicatori finanziari o dati economici, che possono portare a decisioni di credito distorte, a una gestione errata delle frodi e a potenziali perdite finanziarie.

Problemi di fiducia e trasparenza

«solo l'8% delle persone si fiderebbe dei consigli sui mutui offerti da un programma di intelligenza artificiale, una percentuale inferiore al 9% che si fiderebbe del proprio oroscopo per i consigli sugli investimenti (Trust in technology, 2017)»

Fattori determinanti

- Impatto della decisione
- Assenza del decisore umano (accountability)
- Problemi di trasparenza (black box)
- Le persone non mostrano fiducia nei decisori AI
- Le persone tendono a perdere fiducia in un agente AI che commette un errore in misura maggiore rispetto a un umano che commette un errore

AI ethics

Ethics of AI

Ethical AI

Machine Ethics

The autonomy principle

Ethical IA: è possibile?

Autonomia-Responsabilità

Responsabilità-Fiducia

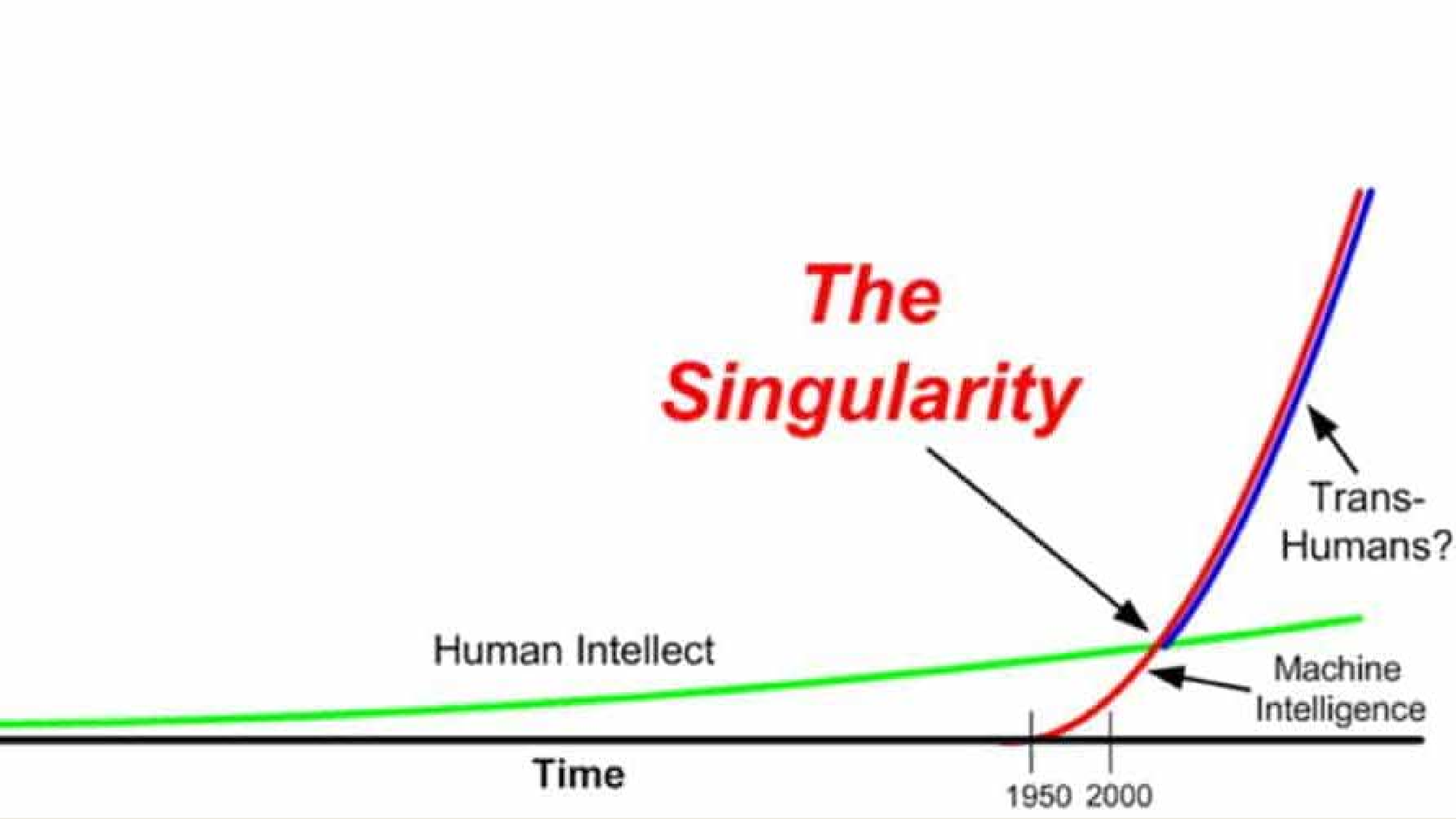
L’IA antropocentrica dell’UE

L’UE nel AI ACT esprime una sostanziale sfiducia nelle applicazioni AI ad elevato livello di autonomia. Le applicazioni che potrebbero favorire eccessi di controllo e manipolazione sociale sono vietate.

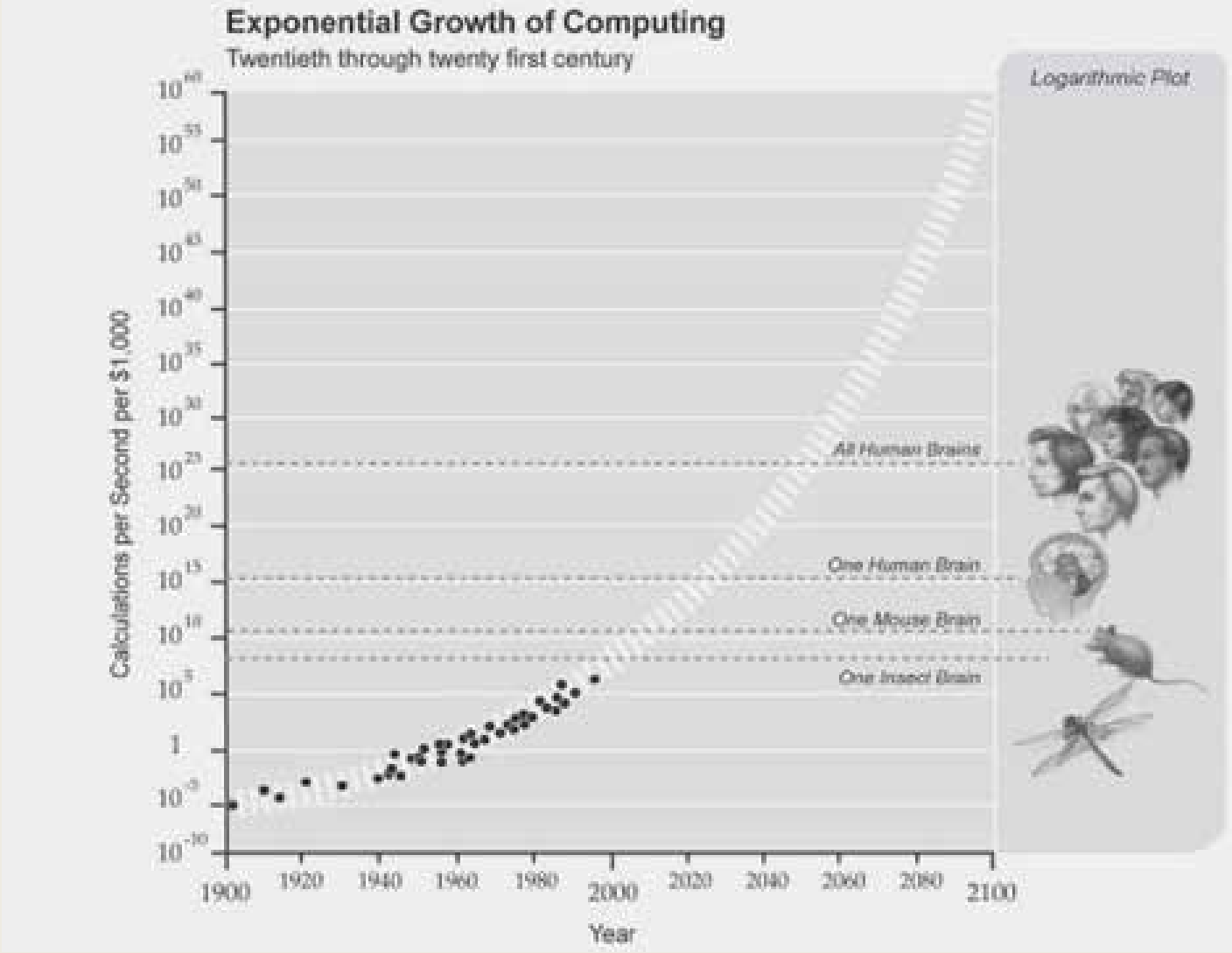
L’AI ACT sembra così rispecchiare le preoccupazioni già discusse in merito a funzionamento non affidabile, sicurezza, bias e opacità, che possono mettere a rischio gli individui e la stessa tenuta democratica.

In questo modo l’UE sembra porsi in controtendenza rispetto a ricerca e mercato, che invece spingono sempre più verso autonomia e autoconsapevolezza dell’IA.

Le tempistiche di attuazione e modifica dell’esistente sembrano un aspetto molto critico, tanto da poter creare una totale inefficacia della regolamentazione.


The Singularity
Trans-Humans?
Human Intellect
Machine Intelligence
Time
1950
2000

The New York Times



Azioni possibili

assicurarsi che i sistemi di IA siano addestrati su dataset rappresentativi e diversificati; condurre verifiche regolari per individuare e correggere eventuali bias

migliorare la trasparenza, creando fiducia tra le parti (explainability, accountability, transparency)

investire in cybersecurity per proteggere i sistemi di IA da potenziali attacchi, anche in vista della prossima rivoluzione quantistica

valutare i rischi attuali, prepararsi alle modifiche normative e comunicare efficacemente con i principali stakeholder per garantire l'allineamento e la comprensione dell'utilizzo dell’IA

intensificare la ricerca su percezioni e atteggiamenti laypeople, credenze in merito a autonomia IA, dilemmi etici, assisted decision-making

Thank you for your attention!

Scan the QR code to learn more about my research.

francesco.labarbera@unina.it