UNIVERSITÀ
POLITECNICA
DELLE MARCHE

DII
Dipartimento di Ingegneria
dell'Informazione

unIMC

Etica e IA: Ethics by design

Jacopo Iezzi

Martedi 19 Settembre 2023

Corso di Perfezionamento in Cybersecurity, Cyber Risk and Data Protection


Frozen Yoghurt
Contains 20% fat


Frozen Yoghurt
80% fat free


E' inutile che parli. Avete già deciso che sono colpevole!
Visto? Sa leggere nel pensiero. E' una strega!

Bias Cognitivo

Il bias cognitivo è un *pattern* sistematico di deviazione dalla norma o dalla razionalità nei processi mentali di giudizio.

I bias cognitivi sono forme di comportamento mentale evoluto: alcuni rappresentano forme di adattamento, in quanto portano ad azioni più efficaci in determinati contesti, o permettono di prendere decisioni più velocemente quando maggiormente necessario

Il *bias* è una forma di distorsione della valutazione causata dal pregiudizio. La mappa mentale di una persona presenta *bias* laddove è condizionata da concetti preesistenti non necessariamente connessi tra loro da legami logici e validi.

Bias e IA

Se la psiche umana ha questo “bug”, è possibile che sia stato propagato nell’IA?

Queste applicazioni si basano su algoritmi che tendono a riflettere (almeno in parte) i preconcetti di chi li ha progettati.

Il pregiudizio può insinuarsi negli algoritmi in diversi modi. In primis, preconcetti, opinioni, aspettative culturali, sociali e istituzionali che preesistono in coloro che ideano e progettano il sistema possono essere trasmessi indirettamente alla tecnologia stessa

Ci sono poi i cosiddetti “bias di incertezza”, un tipo di *bias* che distorce i processi algoritmici verso risultati che riproducono più strettamente i campioni più grandi, ignorando i dati riguardanti popolazioni sottorappresentate nel dataset.

Esempi noti di bias nell'IA

COMPAS Case Study: Investigating Algorithmic Fairness of Predictive Policing

Mallika Chawla · Follow

7 min read · Feb 23, 2022

20

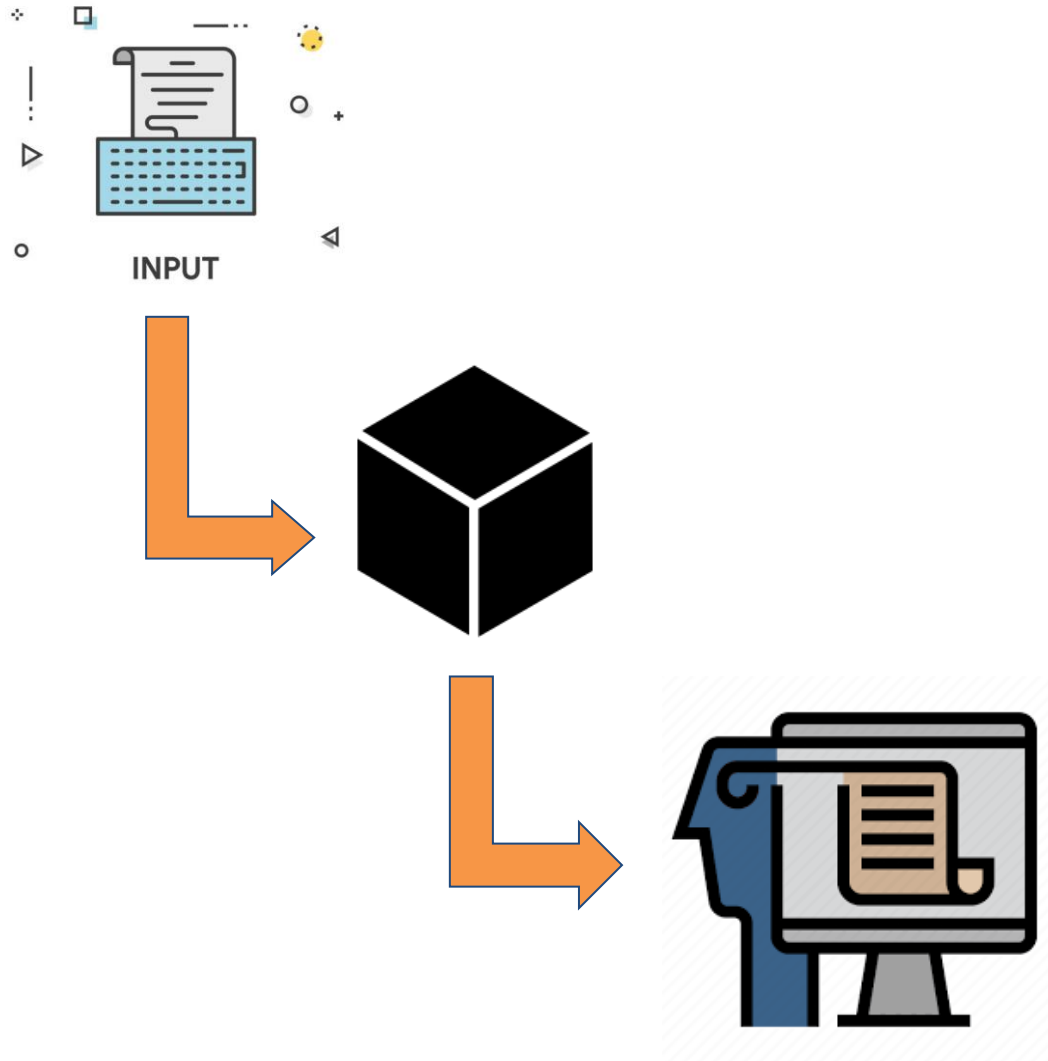
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Healthcare algorithm used across America has dramatic racial biases

System sold by Optum estimates health needs based on medical costs, which are much less than for white patients, report finds

Although the algorithm did not explicitly apply racial identification to patients, it still played out racial biases in effect. Photograph: Andersen Ross/Getty Images/Blend Images

Architettura sistema IA



Immaginiamo il workflow di un sistema di IA

1. Una delle maggiori preoccupazioni nella prima fase è la presenza di errori nel dataset in input.

2. Nella seconda fase, una questione ampiamente dibattuta è la trasparenza e l'accessibilità della procedura - il problema della cosiddetta black box

3. Nella terza fase un aspetto problematico è costituito dai possibili effetti discriminatori della decisione algoritmica.

Guidelines per lo sviluppo di un AI affidabile

IA affidabile

Legalità dell'IA | **Eticità dell'IA** | **Robustezza dell'IA**

(non trattata nel presente documento)

Basi di un'IA affidabile

Garantire l'aderenza ai principi etici basati sui diritti fondamentali

4 Principi etici

Riconoscere e risolvere le potenziali tensioni tra di essi

- Rispetto dell'autonomia umana
- Prevenzione dei danni
- Equità
- Esplicabilità

Realizzazione di un'IA affidabile

Garantire l'attuazione dei requisiti fondamentali

7 Requisiti fondamentali

Da valutare e considerare costantemente durante l'intero ciclo di vita del sistema di IA mediante

- Metodi tecnici
- Metodi non tecnici

- Intervento e sorveglianza umani
- Robustezza tecnica e sicurezza
- Riservatezza e governance dei dati
- Trasparenza
- Diversità, non discriminazione ed equità
- Benessere sociale e ambientale
- Accountability

Valutazione dell'IA affidabile

Garantire l'operatività dei requisiti fondamentali

Valutazione dell'IA affidabile

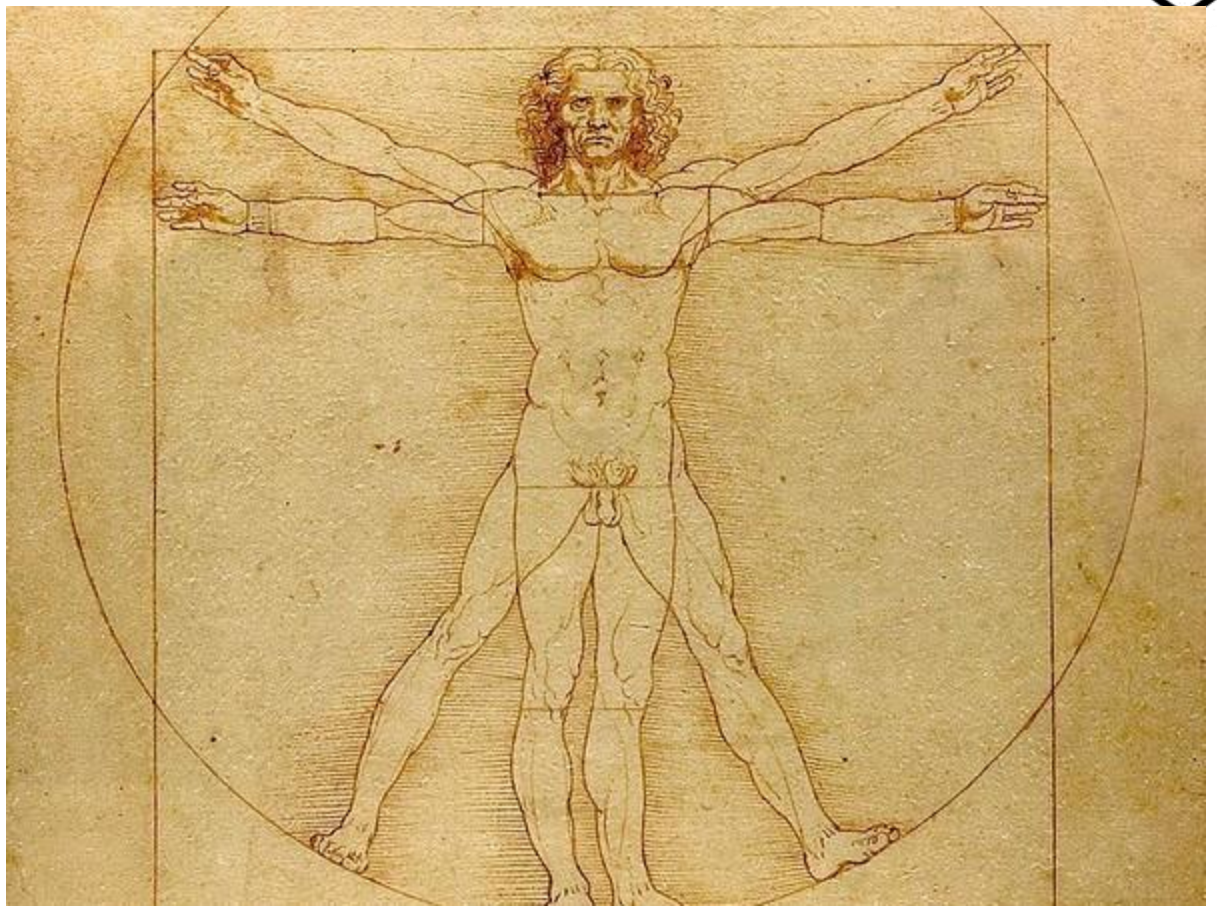
Adattare all'applicazione specifica dell'IA

Principi etici

Principio del rispetto dell'autonomia umana:

diritti fondamentali su cui si fonda l'Unione europea sono volti a garantire il rispetto della libertà e dell'autonomia degli esseri umani.

I sistemi di IA non devono subordinare, costringere, ingannare, manipolare, condizionare o aggregare in modo ingiustificato gli esseri umani



Fonte: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Principi etici

Principio della prevenzione dei danni:

I sistemi di IA non devono causare danni né aggravare e neppure influenzare negativamente gli esseri umani, per cui occorre tutelare la dignità umana nonché l'integrità fisica e psichica.

Fonte: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Principio etici

Equità:

le persone devono avere uguali diritti e opportunità e non devono esserci dei vantaggi o svantaggi .

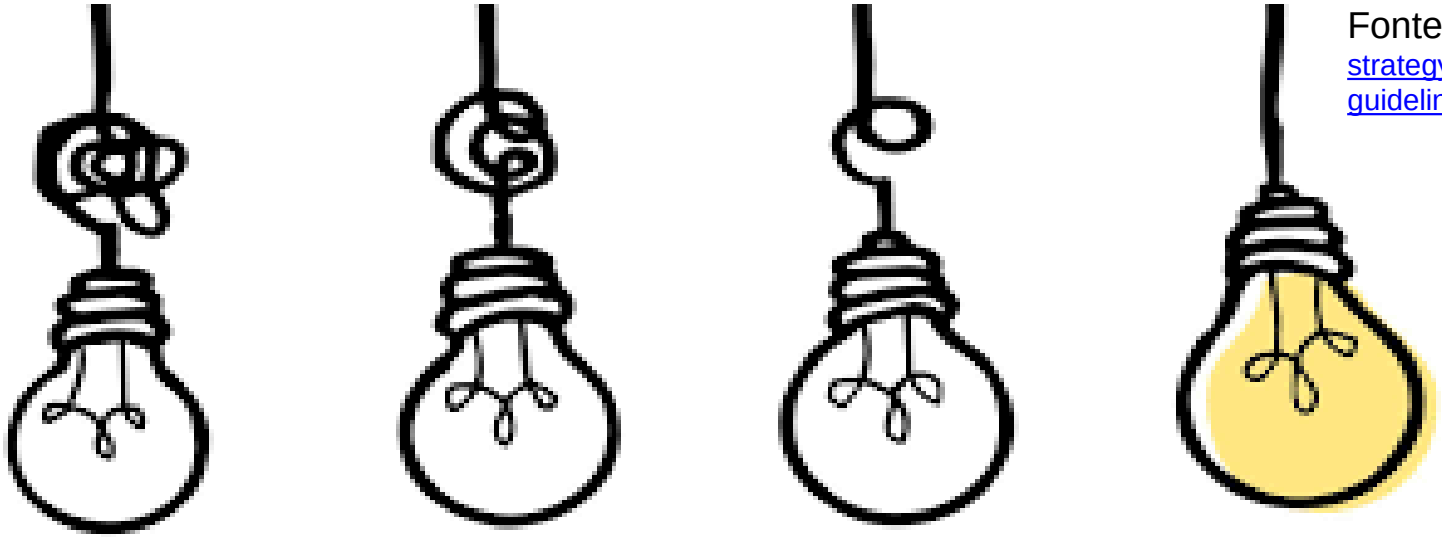
La dimensione sostanziale implica un impegno a garantire una distribuzione giusta ed equa di costi e di benefici e a garantire che gli individui e i gruppi siano liberi da distorsioni inique, discriminazioni e stigmatizzazioni.

Fonte: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Principi etici

Il principio dell'esplicabilità:

processi devono essere trasparenti, le capacità e lo scopo dei sistemi di IA devono essere comunicati apertamente e le decisioni, per quanto possibile, devono poter essere spiegate a coloro che ne sono direttamente o indirettamente interessati



Fonte: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Motivazioni

Come vengono valutati questi principi nel contesto aziendale?

Quali sono i principali vantaggi che spingono le aziende a investire in questo ambito?

Assessment List for Trustworthy AI (ALTAI)

ALTAI è stato sviluppato dal Gruppo di esperti sull'intelligenza artificiale istituito dalla Commissione europea per aiutare a **valutare** se il sistema di IA che viene sviluppato, distribuito, acquistato o utilizzato **è conforme ai sette requisiti dell'IA affidabile**, come specificato nelle nostre *Linee guida etiche per l'IA affidabile.*

Il questionario è suddiviso in 7 categorie. Alla fine delle domande, ALTAI fornisce una misura di quanto l'AI presa in esame sia conforme o meno a queste sette caratteristiche, fornendo suggerimenti su come migliorare il sistema.

Assessment List for prova

Edit Info

Sections of the ALTAI

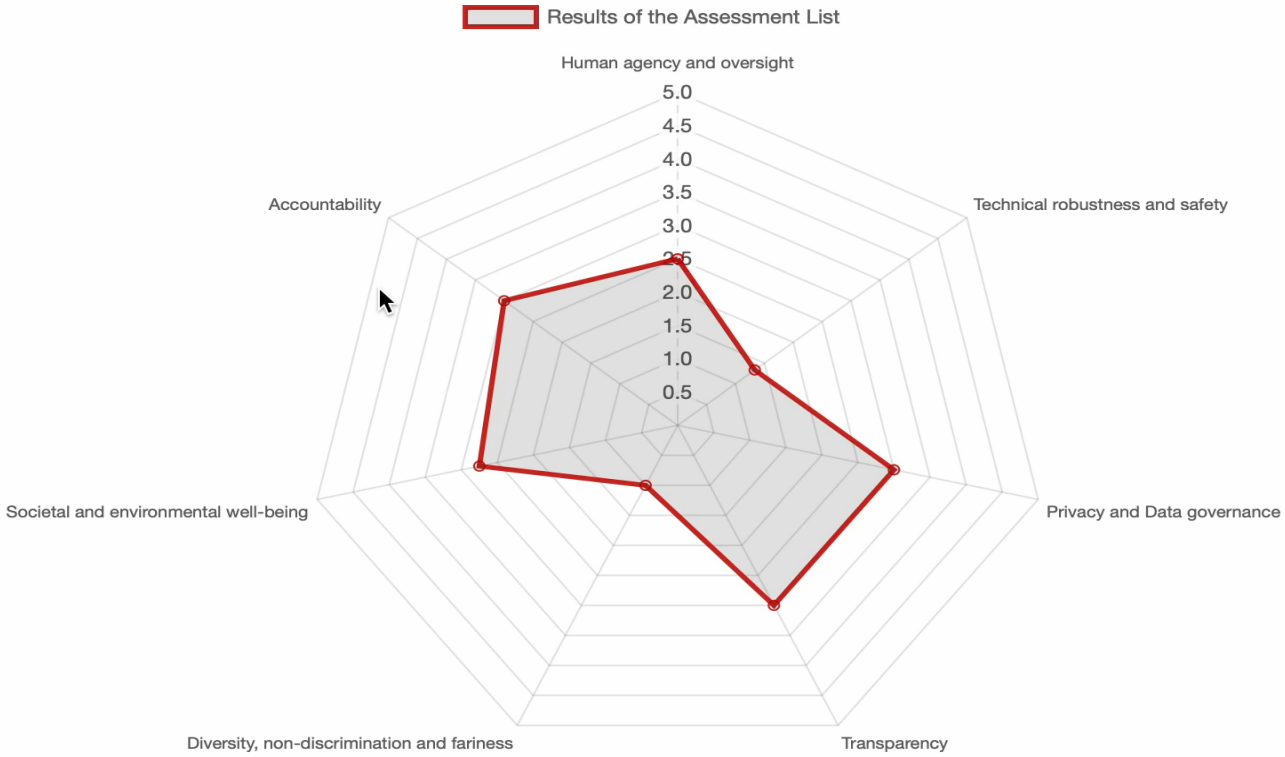
- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-Discrimination and Fairness
- Societal and Environmental Well-being
- Accountability

Legend of progression symbols

- Unanswered
- Partially filled
- Completed and validated

Self assessment results

The requirements not completed score 0.



Recommendations

Human agency and oversight

Ensure that the end-users or subjects are adequately informed that they are interacting with an AI system.

Put in place procedures to avoid that end users over-rely on the AI system.

Principali vantaggi

- **Livello commerciale:** è essenziale comprendere le esigenze legislative quando si sviluppano nuove tecnologie e innovazioni al fine di evitare sanzioni. Inoltre un vantaggio competitivo prevedendo i cambiamenti nelle leggi e nei regolamenti al fine di ottenere una posizione di vantaggio precompetitivo

- **Livello Sociale**: la responsabilità sociale dell'impresa è rientrata a pieno titolo tra le metriche di valutazione delle performance. Permettendo di valutare l'operato dell'azienda non solo in termini economici ma anche in termini di sostenibilità.

Open questions

Si parla molto di IA, principalmente per le paure ad esso connesse.

- Perdita del lavoro
- Perdita di pensiero critico
- Perdita della propria libertà

GRAZIE MILLE

Corso di Perfezionamento in Cybersecurity, Cyber Risk and Data Protection