I pregiudizi della tecnologia: ecco il male nascosto e come curarlo

⋮ 19/6/2024

Nel film “**Brazil**” di Terry Gilliam (1985) c’è la scena iniziale che illustra benissimo quali potrebbero essere **le conseguenze di decisioni fatte affidandosi ad un sistema automatico.** Si vede una stampante che sta producendo i **mandati di cattura** per una serie di criminali considerati terroristi, tra cui un certo **Archibald Tuttle** (interpretato da Robert De Niro).

Ma, a causa della caduta di un insetto tra gli ingranaggi – letteralmente un bug – la stampante sbaglia a scrivere il mandato di cattura che viene erroneamente attribuito a un certo Archibald Buttle, il quale verrà catturato, processato e involontariamente ucciso, così che **un funzionario sarà inviato a rimborsare la vedova per l’errore tecnico.**

Cos’è l’algorithmic bias

Questa scena nonostante l’estetica tecnologica anni ’80, descrive perfettamente quello che potremmo considerare il fenomeno detto **Algorithmic bias** (o anche Machine discrimination: Casillas 2023), ovvero la presenza di pregiudizi e discriminazioni nelle decisioni automatiche

che vengono prese dagli algoritmi – anche quelli di intelligenza artificiale – dovute ai più diversi motivi.

Siamo abituati a pensare che gli algoritmi, in quanto tecnologie, siano necessariamente oggettivi poiché una tecnologia non si lascia influenzare dal contesto di funzionamento: un'automobile non si comporta diversamente a seconda se viene guidata da un uomo o da una donna. In realtà le cose sono un po' più complicate. La tecnologia è un prodotto sociale, e in quanto tale è frutto di un processo sociale che – semplificando – potremmo isolare nelle fasi di ideazione, progettazione, finanziamento, costruzione, commercializzazione, utilizzo, ed ognuna di queste fasi potrebbe essere fonte di fenomeni distorsivi.

Torniamo al nostro esempio dell'automobile: è vero che un'auto è la stessa se viene guidata da un uomo o da una donna, ma potrebbe avere dei **pregiudizi progettuali**, per esempio il sedile del guidatore potrebbe seguire le dimensioni antropometriche degli uomini che – mediamente – sono più alti delle donne; quindi, una donna potrebbe avere difficoltà a sedersi comodamente nel posto di guida. Oppure la disposizione dei finestrini e degli specchietti retrovisori esterni nelle auto ad assetto sportivo potrebbe creare un maggior numero di punti di cieco, che potrebbero rendere l'esperienza di guida più difficoltosa per corpi non conformi: persone mediamente più basse, donne incinte, giovani neo-patentati e così via dicendo. Se questo fenomeno è presente negli artefatti fisici, sicuramente lo è anche negli artefatti immateriali come i software e quindi gli algoritmi.

Il rapporto tra algoritmi e pregiudizi sociali

Esiste una letteratura scientifica piuttosto ampia sul tema del **bias degli algoritmi**, che negli ultimi tempi – complice anche la diffusione dei sistemi di intelligenza artificiale – sta avendo un certo successo poiché sta diventando sempre più necessario sviluppare algoritmi etici, nel senso di non discriminatori e che non provochino dinamiche pregiudiziali di vario tipo.

Ma come fanno gli algoritmi ad esprimere un comportamento discriminatorio se sono modelli formali basati su dati? La letteratura recente ha identificato diversi aspetti attraverso cui un software può esprimere nelle sue decisioni un comportamento pregiudizievole. Per comodità di rassegna metteremo insieme le riflessioni sugli algoritmi in generale e sugli algoritmi di IA (per esempio machine learning) in maniera particolare.

Le fasi di sviluppo degli algoritmi e i possibili pregiudizi

Nelle diverse fasi del loro ciclo di vita, **gli algoritmi possono essere plasmati da tre elementi chiave: i dati, i modelli matematici scelti, il team di sviluppatori** (Ingram 2021, Rejmaniak 2021, Kartal 2022, Fazil *et al.* 2024).

I dati

I dati servono come strumento di **addestramento del sistema**, come modo per testare il funzionamento dell'algoritmo, ma può capitare che i database usati per testarne le caratteristiche possano nascondere dei pregiudizi dovuti alle strategie di raccolta, ai modi con cui sono stati catalogati, alle regole con cui sono stati distribuiti alla comunità scientifica di riferimento (Ntoutsi *et al.* 2020, Ingram 2021, Kartal 2022). Un esempio interessante in questo senso sono gli algoritmi di NLP (Natural Language Processing: elaborazione del linguaggio naturale), i quali essendo allenati su database prodotti da fonti mediali (libri, giornali, riviste) in alcuni casi tendo a riprodurre i pregiudizi taciti o espliciti presenti in tali testi, per esempio associare la parola "programmatore" a "uomo" e la parola "casalinga" a "donna" (Prescott 2023).

I modelli matematici

Anche i modelli matematici possono essere fonte di distorsione: ovviamente non perché la matematica possa essere fonte di distorsione in sé, la matematica è un linguaggio formale e in quanto tale non si presta a produrre bias, però le decisioni che il software deve prendere sulla base del modello matematico possono essere pregiudiziali: per esempio nel caso della valutazione dei dati che sembrano essere anomali (*outlier*), oppure quando bisogna decidere le condizioni in base alle quali un dato deve essere considerato un errore (definizione del rumore) (Ingram 2021, Kartal 2022). Un caso interessante in questo senso è quello di Google translate, ovvero l'algoritmo di traduzione di Google, il quale associava specifici ruolo professionali – professionista, imprenditore, medico – preferibilmente a soggetti di genere maschile invece che a soggetti di genere femminile (Kartal 2022).

Il team di sviluppatori

Per quanto concerne il team di sviluppatori, esso è responsabile della progettazione del sistema decisionale basato su dati; quindi, il sistema può incorporare gli stessi pregiudizi (involontari) del team di progettazione (Ingram 2021, Kartal 2022), oppure del sistema sociale che funge da contesto sovra- o sotto-rappresentando specifici gruppi sociali (Ntoutsi *et al.* 2020). Per esempio **se nel team di progettazione il genere femminile non è presente, è facile che algoritmi pensati per decidere su temi femminili (per esempio legati alla salute) potrebbero non essere funzionali.** Questa situazione porta con sé la necessità che il gruppo che sviluppa i sistemi decisionali automatici deve essere composto nel modo più eterogeneo possibile, così da incorporare nel processo algoritmico anche quelle identità sociali minoritarie o meno presenti nel contesto sociale. Il caso classico per esprimere questa situazione è quello dell'algoritmo di assunzione di Amazon, il quale risultava essere discriminatorio nei confronti delle donne rispetto agli uomini (Prescott 2023). Un altro caso discusso in letteratura è quello dei sistemi di riconoscimento vocale, i quali

hanno delle basse performance per gli accenti non standard e per alcune varietà linguistiche come l'*African American Vernacular English* (Prescott 2023).

Il modello di business come fonte di distorsione

C'è un ulteriore tassello da prendere in considerazione che potrebbe essere fonte di distorsione ed è il modello di business alla base dell'utilizzo del sistema decisionale. **Il modello di business più utilizzato come strumento di raccolta dati è il capitalismo della sorveglianza** (Zuboff 2019), ovvero quella forma economica basata sulla raccolta indiscriminata di dati e il loro uso per alimentare meccanismi commerciali o processi algoritmici. Il problema di questo sistema non è solo legato alla privacy, ma è anche dovuto al fatto che gli algoritmi funzionano sulla base dei dati raccolti attraverso le piattaforme che a loro volta inducono un preciso comportamento sociale che produce gli stessi dati che vengono raccolti per far funzionare gli algoritmi, in un perverso meccanismo autoalimentante che potrebbero portare alla radicalizzazione di diversi pregiudizi (Ingram 2021, Prescott 2023).

Se l'algoritmo di una piattaforma generica – poniamo Instagram – è sensibile all'uso dell'hashtag #foodporn io sarò vincolato ad usare questo specifico hashtag per rendere i miei contenuti maggiormente visibili, che saranno premiati dall'algoritmo di visibilità stesso che rileverà il successo di quello specifico hashtag in un meccanismo che dovrebbe essere neutrale, ma che in realtà è un circolo vizioso.

Gli impatti sociali dell'algorithmic bias

I bias dei sistemi decisionali automatici hanno conseguenze importanti non semplicemente sugli individui, ma sulla società nel suo complesso. In primo luogo perché il loro utilizzo indiscriminato può avere effetto su numero molto grande di persone, inoltre perché gli impatti possono avere conseguenze su ambiti diversi. In un recente studio che ha analizzato quarantacinque diverse ricerche nel periodo 2021-2023 attraverso la tecnica della revisione sistematica, ha identificato quattro diverse categorie principali in cui è possibile classificare gli impatti del bias dell'*artificial intelligence*: diritti fondamentali (discriminazione, disuguaglianza economica/sociale, diritto all'uguaglianza di genere, diritto a un processo equo, amministrazione equa e privacy), individui e società (discriminazione sociale, violazione delle libertà civili, stress psicologico/emotivo e perdita di opportunità), settore finanziario/economico (discriminazione creditizia, differenziazione dei prezzi e accesso ai servizi, e pubblicità finanziaria discriminatoria), imprese e organizzazioni (perdita di fiducia, danni reputazionali e perdite finanziarie) (Bansal *et al.* 2023).

Il caso COMPAS

Un caso emblematico che racchiude quasi tutte le istanze finora analizzate è senza dubbio il caso COMPAS (Rejmaniak 2021, Kartal 2022, Bansal *et al.* 2023, Casillas 2023, Prescott 2023). Con questo acronimo (*Correctional Offender Management Profiling for Alternative Sanctions*), si identifica un particolare algoritmo decisionale usato per valutare il rischio di recidiva criminale e adottato nel sistema giudiziario statunitense. **COMPAS utilizza un insieme di dati storici e caratteristiche personali degli imputati**, come il loro background criminale, età e genere, per fare previsioni sul rischio di recidiva. Gli imputati rispondono a un questionario e i dati vengono elaborati dall'algoritmo per generare un punteggio di rischio. Uno studio condotto da ProPublica (Angwin *et al.* 2016) ha scoperto che il sistema tendeva a sovrastimare il rischio di recidiva per gli imputati neri e a sottostimare il rischio per quelli bianchi. Questo bias era evidente nei falsi positivi (dove gli imputati neri erano erroneamente classificati come ad alto rischio) e nei falsi negativi (dove gli imputati bianchi erano erroneamente classificati come a basso rischio). In pratica **il sistema decideva in maniera razzista**. Le previsioni distorte di COMPAS hanno avuto gravi implicazioni per la vita degli imputati, influenzando decisioni critiche come il rilascio su cauzione, la sentenza e la libertà condizionale. Un bias sistematico contro un gruppo razziale può perpetuare le disuguaglianze e la discriminazione nel sistema giudiziario. Un altro problema è la mancanza di trasparenza. Il codice sorgente e l'algoritmo utilizzato per calcolare i punteggi di rischio sono proprietà di Northpointe (ora Equivant), la società che ha sviluppato COMPAS, e non sono pubblicamente disponibili. Questo rende difficile per i ricercatori e i difensori dei diritti civili analizzare e comprendere come e perché vengono fatte determinate previsioni.

Conclusioni

La società del XXI secolo ha visto profilarsi al suo orizzonte una nuova genìa di soggetti sociali, gli algoritmi decisionali, che saranno sempre più presenti nel nostro presente e nel nostro futuro. È necessario non solo essere consapevoli dell'esistenza di queste distorsioni, ma innescare anche un meccanismo di difesa da tali pregiudizi, sia per evitare di accettare l'importante contributo che i sistemi di intelligenza artificiale daranno alla nostra vita sociale e professionale in maniera acritica, sia per migliorare il nostro rapporto con essi. Perché se è relativamente facile trovare un errore di battitura – da Tuttle a Buttle – è profondamente più difficile dubitare di una decisione se è seppellita in un sistema di reti neurali.

Bibliografia

Angwin, J., Larson, J., Mattu S., Kirchner, L. (2016), *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*, "ProPublica", May 23, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bansal, C., Pandey, K. K., Goel, R., Sharma, A., Jangirala, S. (2023), *Artificial intelligence (AI) bias impacts: classification framework for effective mitigation*, “Issues in Information Systems”, 24(4), 367-389.

Casillas, J. (2023), *Bias and Discrimination in Machine Decision-Making Systems*, In Lara, F., Deckers, J. (eds), *Ethics of Artificial Intelligence*, The International Library of Ethics, Law and Technology, vol 41. Springer.

Fazil, A. W., Hakimi, M., Shahidzay, A. K. (2023), *A Comprehensive Review Of Bias In Ai Algorithms*, “Nusantara Hasana Journal”, 3(8), 1-11.

Ingram, K. (2020), *Constructing AI: Examining how AI is shaped by data, models and people*, “The International Review of Information Ethics”, 29(03).

Kartal, E. (2022), *A Comprehensive Study on Bias in Artificial Intelligence Systems: Biased or Unbiased AI, That’s the Question!*, “International Journal of Intelligent Information Technologies” (IJIIT), 18(1), 1-23.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., … Staab, S. (2020), *Bias in data‐driven artificial intelligence systems. An introductory survey*, “Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery”, 10(3), e1356.

Prescott, A. (2023), *Bias in Big Data, Machine Learning and AI: What Lessons for the Digital Humanities?*, “DHQ: Digital Humanities Quarterly”, 17(2).

Rejmaniak, R. (2021), *Bias in Artificial Intelligence Systems*, “Białostockie Studia Prawnicze”, 3(26), 25-42.

Zuboff, S. (2019), *Il capitalismo della sorveglianza*, Luiss University Press, Roma 2020.

@RIPRODUZIONE RISERVATA