7. La mappatura dell'etica degli algoritmi

Una definizione operativa di algoritmo

Definizione di algoritmo rilevante in questo ambito:

> **costrutto matematico**, con "*una struttura di controllo finita, astratta, efficace, composta, data in modo imperativo, che realizza un dato scopo sotto determinate condizioni*"

Ci concentriamo sulle questioni etiche poste dagli algoritmi come costrutti matematici, dalle loro implementazioni come programmi e configurazioni (applicazioni) e dai modi in cui tali questioni possono essere affrontate.

È risaputo che gli algoritmi non sono eticamente neutri:

```
- i risultati degli algoritmi di traduzione e dei motori di ricerca siano
largamente percepiti quali oggettivi, anche se spesso codfiicano il linguaggio
con modalità condizionate dal genere
- la presenza di pregiudizi è stata ampiamente segnalata, per esempio nella
pubblicità algoritmica, con opportunità di lavori più remunerativi e di
impieghi nel campo della scienza e della tecnologia pubblicizzati più spesso
per gli uomini che per le donne
- gli algoritmi di previsione utilizzati per gestire i dati sanitari di
milioni di pazienti negli Stati Uniti aggravano i problemi esistenti, con
pazienti bianchi che ricevono cure signicativamente migliori rispetto a
pazienti di colore che si trovano in situazioni analoghe
```

Oggi l'AI sta attraversando una nuova "estate", sia per i progressi tecnici in atto sia per l'attenzione che il settore ha ricevuto; vi è stato quindi un incremento considerevole delle ricerche sulle implicazioni etiche degli algoritmi, in particolare in relazione agli aspetti di **equità, responsabilità e trasparenza**.

La mappa etica degli algoritmi

> Si possono usare algoritmi
>
> 1. per trasformare i dati in prove (informazioni) per un dato risultato che si usa
> 2. per innescare e motivare un'azione che può avere conseguenze etiche.
>
> Le azioni (1) e (2) possono essere eseguite da algoritmi (semi) autonomi, come gli algoritmi di apprendimento automatico (ML), e questo complica una terza azione, vale a dire:
>
> 3. attribuire la responsabilità degli effetti delle azioni che un algoritmo può innescare.

Nel contesto di (1)-(3), il *Deep Learning* è di particolare interesse, in quanto campo che include architetture di deep learning (apprendimento profondo).

I sistemi informatici che implementano algoritmi di possono essere descritti come “autonomi” o “semi-autonomi”, nella misura in cui i loro risultati sono indotti dai dati e quindi non sono deterministici.

In base a questo approccio, si identificano, con la mappa concettuale riportata sotto, **sei questioni etiche**, che definiscono lo spazio concettuale dell’etica degli algoritmi in quanto ambito di ricerca.

Tre delle questioni etiche si riferiscono a fattori epistemici, in particolare: prove **inconcludenti, imperscrutabili e fuorvianti**;
due sono esplicitamente normative: **esiti ingiusti ed effetti trasformativi;**
mentre una, la **tracciabilità**, è rilevante a fini sia epistemici sia normativi.

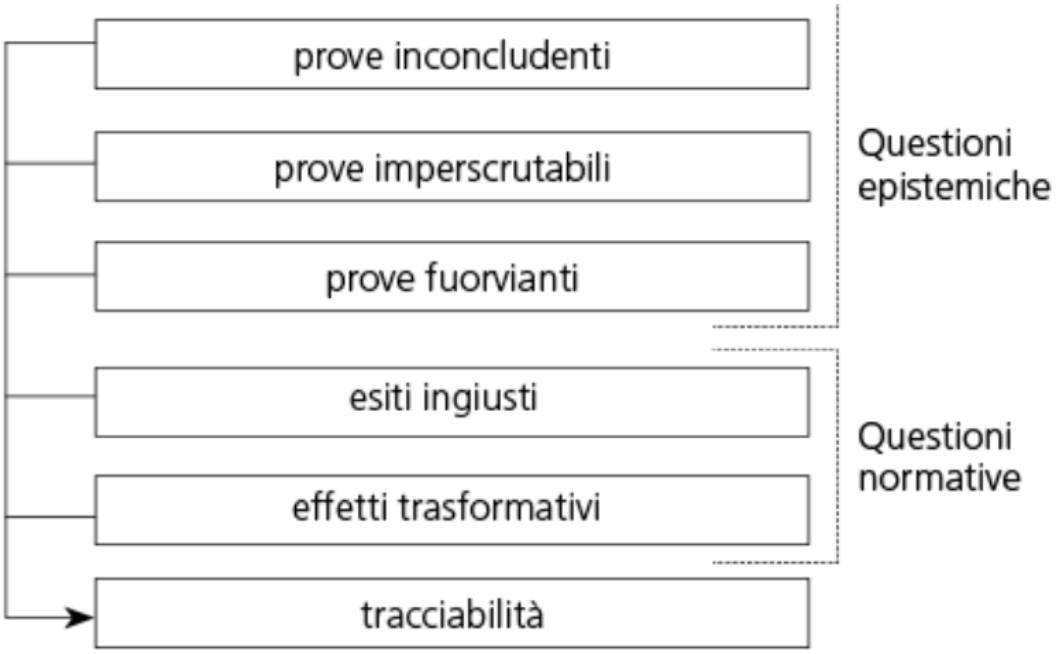

Figura 7.1 Sei tipi di questioni etiche sollevate dagli algoritmi (Mittelstadt, Allo, Taddeo et al., 2016, p. 4).

1. Prove inconcludenti che portano ad azioni ingiustificate

> La ricerca incentrata su prove inconcludenti si riferisce al modo in cui gli algoritmi non deterministici di producono output espressi in termini probabilistici
> Questi tipi di algoritmi generalmente identicano l’associazione e la correlazione tra le variabili nei dati sottostanti, ma non le connessioni causali. In quanto tali, possono incoraggiare la pratica dell’apofenia:
> vedere schemi ricorrenti (patterns) dove in realtà non ne esistono, semplicemente perché enormi quantità di dati possono offrire connessioni che si irradiano in tutte le direzioni

Ricerche recenti hanno rimarcato la preoccupazione che prove inconcludenti possano dar luogo a gravi rischi etici.
Per esempio, concentrarsi su indicatori non causali può distogliere l’attenzione dalle cause alla base di un determinato problema

Anche con l’uso di metodi causali, i dati disponibili potrebbero non contenere sempre informazioni sufficienti per giusticare un’azione o rendere equa una decisione.

Infatti, le informazioni che possono essere estratte dai dati dipendono fortemente dai presupposti che hanno guidato il processo di raccolta dei dati.

Il **rischio più grande delle prove inconcludenti** è che queste vengano considerate attendibili e che, sulla base di questa assunzione, vengano delegate scelte e responsabilità al processo automatico, a sfavore del sapere umano derivato dall’esperienza.

> Questo è il motivo per cui è fondamentale garantire che i dati forniti agli algoritmi siano **convalidati in modo indipendente** e che siano messe in atto **misure di conservazione e riproducibilità dei dati** per mitigare le prove inconcludenti che portano ad azioni ingiusticate, insieme a processi di **auditing** per identificare risultati ingiusti e conseguenze non volute

2. Prove imperscrutabili che portano all'opacità

> Le **prove imperscrutabili** riguardano i problemi legati alla **mancanza di trasparenza** che spesso caratterizzano gli algoritmi, in particolare algoritmi e modelli di AI, l’infrastruttura sociotecnica in cui essi esistono e le decisioni che supportano

L’assenza di trasparenza – intrinsecamente dovuta ai limiti della tecnologia oppure dovuta a vincoli giuridici in termini di proprietà intellettuale – si traduce spesso in una **mancanza di controllo e/o di responsabilità**.

Secondo studi recenti, i fattori che contribuiscono alla mancanza generale di trasparenza algoritmica includono:

- l’impossibilità cognitiva per gli esseri umani di interpretare giganteschi modelli algoritmici e insiemi di dati;
- una mancanza di strumenti appropriati per visualizzare e tenere traccia di grandi volumi di codice e dati;
- codice e dati così mal strutturati da essere impossibili da leggere;
- aggiornamenti continui e influenza umana sul modello

È importante sottolineare che, se certamente è reale la difficoltà di spiegare l’output degli algoritmi di AI, è al contempo importante non lasciare che questa difficoltà **incentivi le organizzazioni a sviluppare sistemi complessi per sottrarsi alle responsabilità**.

> La trasparenza non è un principio etico in sé, ma una condizione pro-etica per consentire o frenare altre pratiche o principi etici.

Talora, l’opacità può essere più utile, per esempio, per assicurare la segretezza delle preferenze e dei voti politici dei cittadini, o per garantire la concorrenza nelle aste per i servizi pubblici. Infatti, anche in contesti algoritmici, la completa trasparenza può causare essa stessa specifici problemi etici:
- Può fornire agli utenti informazioni rilevanti sulle caratteristiche e sui limiti di un algoritmo
- Può anche sovraccaricare gli utenti di informazioni e in tal modo rendere l’algoritmo più opaco

Esistono diversi modi per affrontare i problemi legati alla mancanza di trasparenza:

> ```
> ogni componente, non importa quanto semplice o complesso, deve essere accompagnato da una scheda tecnica che ne descrive le caratteristiche operative, i risultati dei test, l’utilizzo consigliato e altre informazioni.
> ```

oppure

> ```
> utilizzo di strumenti tecnici per testare e controllare i sistemi algoritmici e il processo decisionale
> ```

oppure

> ```
> considerare i "fattori di trasparenza" attraverso quattro livelli di sistemi algoritmici:
> - dati
> - modello
> - inferenza
> - interfaccia
> ```

La spiegabilità è particolarmente importante se si considera il numero in rapida crescita di modelli e insiemi di dati open source e di facile utilizzo.
Ciò ha spinto gli studiosi a suggerire che, per affrontare il problema della complessità tecnica, è necessario **investire maggiormente nell’istruzione pubblica** per migliorare l’**alfabetizzazione computazionale e relativa ai dati**.

3. Prove fuorvianti che portano a pregiudizi (BIAS) non voluti

Alcuni studiosi si riferiscono al pensiero dominante nel campo dello sviluppo di algoritmi nei termini di “*formalismo algoritmico*”, caratterizzato dall’adesione a regole e forme prescritte.

Sebbene questo approccio sia utile per astrarre e denire i processi analitici, tende a ignorare la complessità sociale del mondo reale.

Alcuni studiosi sottolineano i limiti delle astrazioni per quanto riguarda i pregiudizi non voluti negli algoritmi e sostengono la necessità di sviluppare una cornice sociotecnica per arontare e migliorare l’equità degli algoritmi.
A questo proposito, indicano cinque “trappole” dell’astrazione, o incapacità di rendere conto del contesto sociale in cui operano gli algoritmi, che permangono nel design algoritmico a causa dell’assenza di una cornice sociotecnica, vale a dire:

1. l’incapacità di **modellare l’intero sistema a cui sarà applicato un criterio sociale**, come l’equità;
2. l’incapacità di comprendere come la riproposizione di soluzioni algoritmiche disegnate per un contesto sociale possa risultare fuorviante, imprecisa o comunque arrecare danno se

applicata a un **contesto diverso**;

3. l'**incapacità di rendere** pienamente **conto del signicato di concetti sociali come equità**, che possono essere procedurali, contestuali e discutibili, e non possono essere risolti tramite formalismi matematici;
4. l'incapacità di comprendere come l'**introduzione di una tecnologia in un sistema sociale esistente modichi i comportamenti e i valori incorporati nel sistema preesistente**;
5. l'incapacità di riconoscere che la migliore **soluzione a un problema possa non coinvolgere la tecnologia**.

> Il termine "pregiudizio" (**bias**) ha spesso un'accezione negativa, ma qui è usato per indicare una "deviazione da uno standard", che può vericarsi in qualsiasi fase del processo di design, sviluppo e implementazione.
>
> I **dati** utilizzati per addestrare un algoritmo sono una delle principali **fonti da cui emerge il pregiudizio**, attraverso dati campionati in modo preferenziale o da dati che riettono pregiudizi sociali già **esistenti**

Un possibile approccio per mitigare questo problema consiste nell'**escludere intenzionalmente alcune speciche variabili di dati dalla formazione del processo decisionale algoritmico**.
In effetti, il trattamento di variabili sensibili statisticamente rilevanti o di "variabili protette", come il genere o la razza, è tipicamente limitato o vietato dal diritto antidiscriminatorio e dalla protezione dei dati, al fine di limitare i rischi di sleale discriminazione.

Purtroppo, anche se la protezione di speciche classi può essere codicata in un algoritmo, potrebbero sempre esserci:

- dei pregiudizi (bias) che non sono stati considerati **ex ante**, come nel caso, per esempio, di modelli linguistici che riproducono testi fortemente maschilisti.
- i **proxy non previsti** per queste variabili potrebbero essere comunque usati per ricostruire i pregiudizi, portando a "pregiudizi basati su proxy
    - ad esempio, pregiudizi relativi al codice postale

Approcci più semplici per mitigare la distorsione nei dati comportano:

- la gestione di algoritmi in diversi contesti e con vari insiemi di dati
- rendere pubblico un modello, i suoi insiemi di dati e i metadati (sulla provenienza), al ne di consentire un controllo esterno, può contribuire a correggere pregiudizi invisibili o indesiderati
- generazione di dati equi (ad esempio attraverso reti GANs)

4. Risultati ingiusti che portano alla discriminazione

Ci sono numerose sfumature nella definizione, stima e applicazione di diversi standard di equità algoritmica.
Per esempio, l'equità algoritmica può essere definita in relazione **sia a gruppi sia a individui**.

Per queste e altre ragioni correlate, di recente hanno acquisito importanza quattro definizioni principali di equità algoritmica:

1. **Anti-classificazione**: che fa riferimento a categorie protette, come razza e genere, e i loro proxy utilizzati in modo implicito nel processo decisionale;
2. **parità di classicazione**, che considera un modello equo se le misurazioni comuni delle prestazioni predittive, inclusi i tassi di falsi positivi e negativi, sono uguali tra i gruppi protetti;
3. **calibrazione**, che considera l'equità come una misura di quanto sia ben calibrato un algoritmo tra gruppi protetti;
4. **parità statistica**, che denisce l'equità come una stima uguale di probabilità media relativa a tutti i membri dei gruppi protetti.

Tuttavia, ciascuna di queste denizioni comunemente utilizzate di equità presenta degli svantaggi; inoltre, sono in genere reciprocamente incompatibili.

```
Prendendo per esempio l'anti-classicazione, le caratteristiche protette, come
razza, genere e religione, non possono essere semplicemente rimosse dai dati
di addestramento per prevenire la discriminazione, come osservato sopra.
Le disuguaglianze strutturali signicano che punti dati formalmente non
discriminatori, come i codici postali, possono fungere da proxy ed essere
utilizzati, intenzionalmente o no, per inferire caratteristiche protette, come
la razza.

Inoltre, ci sono casi rilevanti in cui è opportuno considerare le
caratteristiche protette per prendere decisioni eque. Per esempio, tassi di
recidiva femminile più bassi signicano che l'esclusione del genere come input
negli algoritmi di recidiva comporterebbe per le donne valutazioni di rischio
sproporzionatamente elevate
```

Per questo motivo, è importante **considerare il contesto storico e sociologico**, che può modellare approcci appropriati dal punto di vista contestuale all'equità negli algoritmi.

Per quanto riguarda i metodi per migliorare l'equità algoritmica si propongono 2 approcci:

- l'intervento **di una terza parte** che disponga di dati su caratteristiche sensibili o protette e tenti di identicare e ridurre le discriminazioni causate dai dati e dai modelli
- un metodo collaborativo basato sulla conoscenza che si concentri su risorse di dati generate dalla comunità che comprendano esperienze pratiche di e modellazione

5. Effetti trasformativi che sollevano sfide per l'autonomia e la privacy informativa

L'impatto collettivo degli algoritmi ha stimolato discussioni sull'autonomia accordata agli utenti finali.

I limiti all'autonomia degli utenti derivano da tre fonti:

- la distribuzione pervasiva e la proattività degli algoritmi (di apprendimento) nel modellare le scelte degli utenti
- la comprensione limitata degli algoritmi da parte degli utenti;
- la mancanza di potere di secondo ordine (o di appelli) nei confronti dei risultati algoritmici

L'autonomia umana può anche essere limitata dall'incapacità di un individuo di comprendere alcune informazioni o di prendere le decisioni appropriate.
Una questione chiave individuata nei dibattiti sull'autonomia degli utenti è la difficoltà di trovare un giusto **equilibrio tra il processo decisionale delle persone e quello delegato agli algoritmi**.

La **privacy informativa** è intimamente legata all'autonomia degli utenti.

> La privacy informativa garantisce la libertà degli individui di pensare, comunicare e formare relazioni, tra le altre attività umane essenziali.
> Tuttavia, la crescente interazione degli individui con i sistemi algoritmici ha eettivamente **ridotto la loro capacità di controllare** chi ha accesso alle informazioni che li riguardano e che cosa viene fatto con tali informazioni.
>
> Pertanto, le grandi quantità di dati sensibili richiesti nella prolazione e nelle previsioni algoritmiche, fondamentali per i sistemi di raccomandazione, sollevano molteplici problemi al riguardo della privacy informativa degli individui.
>
> In effetti, la profilazione algoritmica si basa anche su informazioni raccolte su **altri individui e gruppi di persone che sono stati classfiicati in modo simile alla persona oggetto della profilazione**.
>
> Sebbene ciò ponga un problema di **prove inconcludenti**, indica anche che, se non viene assicurata la **privacy di gruppo**, può risultare impossibile per gli individui sottrarsi al processo di profilazione e predizione algoritmiche.

Gli utenti potrebbero non essere sempre a conoscenza, o avere la capacità di acquisire consapevolezza, del tipo di informazioni che sono detenute al loro riguardo e dell'utilizzo che di tali informazioni viene fatto. Dato che i sistemi di raccomandazione contribuiscono alla costruzione dinamica dell'identità degli individui intervenendo nelle loro scelte, l'assenza di controllo sulle proprie informazioni si traduce in una perdita di autonomia.

Conferire agli individui la possibilità di contribuire al design di un sistema di raccomandazione può contribuire a creare profili più accurati che tengano conto di attributi e categorie sociali che altrimenti non sarebbero stati inclusi nella classfiicazione utilizzata dal sistema per analizzare gli utenti.

Infine, un sapere crescente in tema di **privacy differenziale** sta fornendo nuovi metodi di protezione della privacy per le organizzazioni che cercano di proteggere la privacy dei propri utenti pur mantenendo una buona qualità del modello, nonché costi e complessità del soware gestibili, trovando un equilibrio tra utilità e privacy.

Tracciabilità come presupposto della responsabilità morale

> Le limitazioni tecniche di vari algoritmi di , come la mancanza di trasparenza o di spiegabilità, minano la possibilità di sottoporli a esame ed evidenziano la necessità di nuovi approcci per tracciare la responsabilità morale e per rendere conto delle azioni poste in essere dagli algoritmi di AI

La complessità tecnica e il dinamismo degli algoritmi di li rendono inclini a questioni di "**riciclaggio dell'agire**": un errore morale che consiste nel prendere le distanze da azioni moralmente sospette, indipendentemente dal fatto che tali azioni siano o no intenzionali, dando la colpa all'algoritmo.

Per affrontare questo problema, bisogna istituire **organismi separati per la supervisione etica degli algoritmi**.

> I problemi relativi al "riciclaggio dell'agire" e all' "elusione dell'etica" derivano dall'inadeguatezza delle cornici concettuali esistenti nel tracciare e attribuire la **responsabilità morale**.
>
> Floridi suggerisce di attribuire la piena responsabilità morale "per impostazione predenita e in modo reversibile" a **tutti gli agenti morali** (per esempio, umani o costituiti da esseri umani, come le aziende) nella rete che sono causalmente rilevanti per una data azione della rete.