1. Passato - l'origine dell'intelligenza artificiale

Come gli sviluppi della digitalizzazione hanno creato le condizioni per l'attuale diusione e per il successo dei sistemi di AI?

Una maggiore **potenza di calcolo** e una **maggiore quantità di dati** hanno reso possibile il passaggio dalla logica alla statistica.
Le reti neurali che erano interessanti solo da un punto di vista teorico sono diventate strumenti ordinari nell'ambito dell'apprendimento automatico.
La vecchia era per lo più **simbolica** e poteva essere interpretata come una branca della logica matematica, ma la nuova è principalmente connessionista e potrebbe essere interpretata come una branca della statistica.

> La potenza e la velocità di calcolo, le dimensioni della memoria, la quantità di dati, i potenti algoritmi, gli strumenti statistici e le interazioni online sono tutti fattori che stanno crescendo in modo incredibilmente rapido.
>
> Ciò accade anche perché il **numero di dispositivi digitali** che interagiscono tra loro è già notevolmente superiore alla popolazione umana.
>
> Perciò, la maggior parte delle comunicazioni avviene da macchina a macchina, senza coinvolgimento umano.

Un numero crescente di persone vive sempre più diffusamente **onlife**, sia online sia offline, e nell'**Infosfera**, sia digitalmente sia analogicamente.

Il potere di scissione del digitale - tagliare e incollare la modernità

Le tecnologie, le scienze, le pratiche, i prodotti e i servizi digitali, in breve il digitale come fenomeno globale sta profondamente trasformando la realtà. Tutto questo è piuttosto ovvio e pacifico.

Le vere domande consistono casomai nel chiedersi **perché**, **come** e con quali **conseguenze**.

Il digitale "taglia e incolla" le nostre realtà sia ontologicamente sia epistemologicamente, ovvero incolla, scolla o rincolla certi aspetti del mondo – e quindi le nostre corrispondenti ipotesi su di essi – che pensavamo fossero immutabili.

> *Esempi di incollamento*
>
> La nostra **identità** e i nostri **dati personali** non sono mai stati incollati insieme così indistinguibilmente come accade oggi, allorché si parla di **identità personale dei soggetti interessati**.
>
> **Posizione e presenza**:
> In un mondo digitale, è ovvio che uno può trovarsi sicamente in un posto, diciamo un bar, ed essere presente interattivamente in un altro, diciamo una pagina su Facebook. Eppure tutte le generazioni passate che vivevano in un mondo esclusivamente analogico hanno concepito e sperimentato posizione e presenza come due lati inseparabili della stessa situazione umana: l'essere situati nello spazio e nel tempo, qui e ora.
>
> **Legge e territorialità**: per secoli, la decisione del giudice si applica nel perimetro dei confini nazionali entro cui opera l'autorità della legge.
> Tuttavia, Internet non è uno spazio fisico, e il problema della territorialità si profila a partire dallo scollamento ontologico tra lo spazio normativo del diritto, lo spazio fisico della geografia e lo spazio logico del digitale.
>
> Per esempio, lo scollamento tra legge e territorialità è diventato tanto palese quanto problematico durante il dibattito sul cosiddetto **diritto all'oblio**.
>
> Ulteriori esempi:
>
> - **realtà virtuale** (scollamento) e **realtà aumentata** (incollamento);
> - **uso** e **proprietà** nella share economy;
> - **autenticità e memoria** grazie alla blockchain;
> - *reddito di base universale*, un caso di scollamento tra **stipendio e lavoro**.

Perché il digitale ha questo potere di scissione, vale a dire di incollare, scollare o ri-incollare il mondo?

La risposta, sta nella combinazione di due fattori:

- è una tecnologia di terzo ordine

A causa dell'autonoma potenza di calcolo del digitale, potremmo anche non avere controllo sul (per non parlare di essere parte del) processo.

- non è semplicemente qualcosa che potenzia o aumenta una realtà, ma qualcosa che la trasforma radicalmente, perché crea nuovi ambienti che abitiamo e nuove forme di agire con cui interagiamo.

Etica, governance e design

Il potere di scissione del digitale **riduce enormemente i vincoli della realtà e ne aumenta le possibilità**.
Trarre vantaggio da tali possibilità e vincoli in vista della risoluzione di alcuni problemi è ciò che possiamo denire **design**.

Ogni epoca ha innovato la propria cultura, società e ambiente facendo adamento su almeno tre elementi principali: la **scoperta**, l'**invenzione** e il **design**.

L'età post-rinascimentale e la prima modernità possono essere qualicate come l'epoca delle scoperte, soprattutto geograche;
la tarda modernità è ancora un'epoca di scoperte ma, con le sue innovazioni industriali e meccaniche, è forse in misura maggiore un'epoca di invenzioni;
la nostra epoca in modo peculiare e più di ogni altra è l'età del design.

1.3 Etica, governance e design

Il potere di scissione del digitale **riduce enormemente i vincoli della realtà e ne aumenta le possibilità**.
Trarre vantaggio da tali possibilità e vincoli in vista della risoluzione di alcuni problemi è ciò che possiamo denire **design**.

Ogni epoca ha innovato la propria cultura, società e ambiente facendo adamento su almeno tre elementi principali: la **scoperta**, l'**invenzione** e il **design**.

L'età post-rinascimentale e la prima modernità possono essere qualicate come l'epoca delle scoperte, soprattutto geograche;
la tarda modernità è ancora un'epoca di scoperte ma, con le sue innovazioni industriali e meccaniche, è forse in misura maggiore un'epoca di invenzioni;
la nostra epoca in modo peculiare e più di ogni altra è l'età del design.

1.4 Nuove forme dell'agire

> Il digitale ha cambiato la natura dell’agire, ma stiamo ancora interpretando l’esito di tali cambiamenti attraverso una mentalità moderna, e ciò genera qualche profondo malinteso.

Negli attuali dibattiti sulla democrazia diretta, talora siamo indotti erroneamente a credere che il digitale dovrebbe ricongiungere *sovranità* (il potere politico che può essere legittimamente delegato) e *governance* (il potere politico che è legittimamente delegato, temporaneamente, condizionatamente e responsabilmente, e che può essere in modo altrettanto legittimo ripreso). La democrazia rappresentativa è comunemente (benché erroneamente) concepita come un compromesso dovuto a vincoli pratici di comunicazione.

Eppure questo è un errore, perché la democrazia indiretta è sempre stata il vero progetto da realizzare.
La disgiunzione è una caratteristica e non un difetto, per dirlo in modo esplicito.
E ciò perché un regime democratico è prima di tutto caratterizzato non da talune procedure o da alcuni valori (elementi da cui pure è caratterizzato), ma da una chiara e netta separazione – cioè disgiunzione – tra coloro a cui appartiene il potere politico (sovranità) che delegano legittimamente con il voto (di tutti i cittadini che vi hanno diritto) e coloro a cui è adato questo potere politico (governance) che esercitano in forza di tale mandato, governando in modo trasparente e responsabile, ntanto che vi sono legittimamente autorizzati.

2. Presente - IA come nuova forma dell'agire e non dell'intelligenza

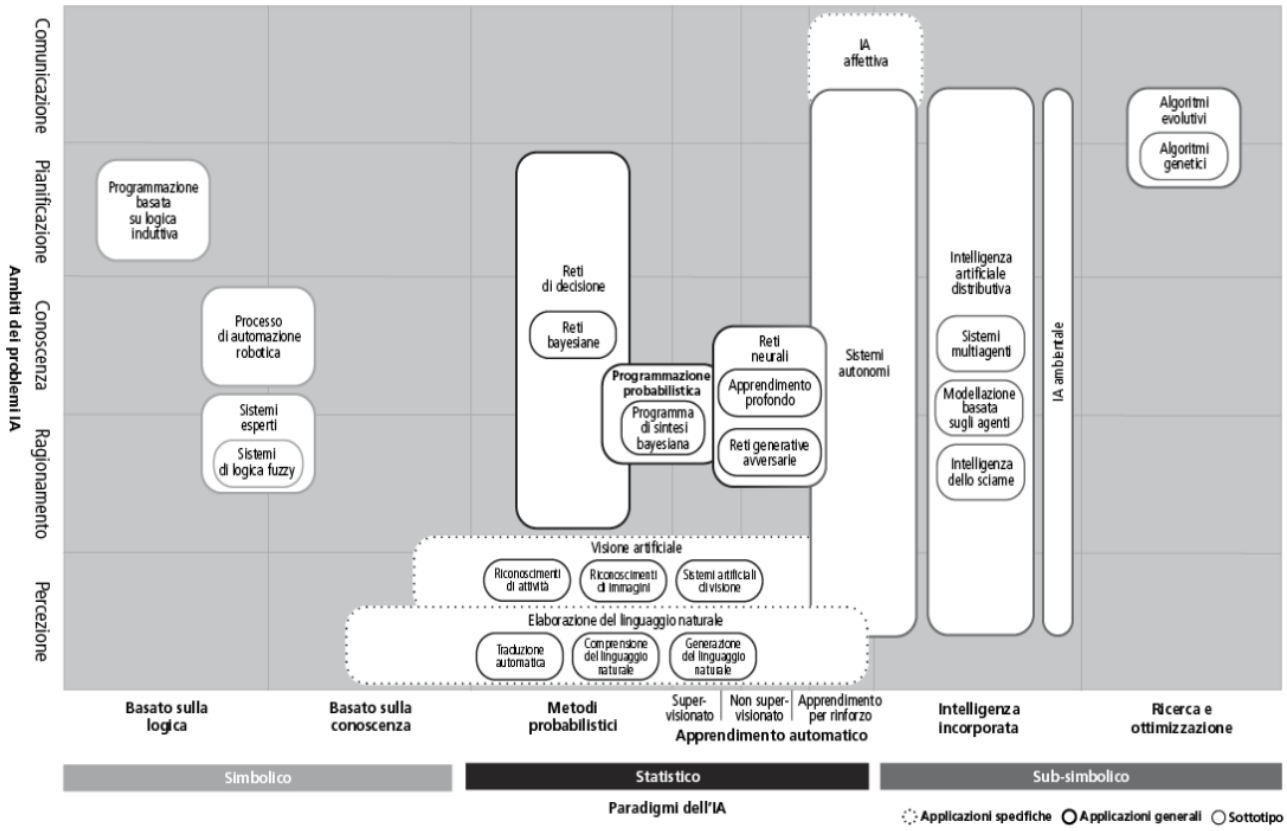
L’IA è stata definita in molti modi, ma non esiste una sua definizione unitaria su cui tutti concordino.

Di fronte a una sfida simile, Wikipedia risolve il problema optando per una tautologia:

> L’intelligenza artificiale è l’intelligenza mostrata dalle macchine, in contrasto con l’intelligenza naturale mostrata dagli esseri umani. (Wikipedia, “Articial Intelligence”, 17 gennaio 2020)

Ciò è al contempo assolutamente vero e totalmente inutile.

L’assenza di una denizione standard di può essere un problema perché, quasi inevitabilmente, in un seminario sull’etica dell’ prima o poi qualche brillante partecipante non può fare a meno di chiedersi, pensieroso: “Ma cosa si intende veramente per IA?”



Mappa della conoscenza IA

Definizione che adotteremo:

> Per il presente scopo il problema dell’intelligenza articiale è quello di far sì che una macchina agisca con modalità che sarebbero denite intelligenti se un essere umano si comportasse allo stesso modo. (Citazione dalla riedizione del 2006 in McCarthy, Minsky, Rochester, Shannon, 2006)

Questa non ha nulla a che vedere con il **pensiero** ma esclusivamente con il **comportamento**: se un essere umano si comportasse in quel modo, quel comportamento sarebbe denito intelligente.
Non signica che la macchina sia intelligente o che addirittura stia **pensando**.

La comprensione controfattuale dell'AI è alla base anche del test di Turing.
Turing comprese molto bene che non vi era modo di rispondere alla domanda se una macchina fosse in grado di pensare, perché, come ammise, entrambi i termini sono privi di definizione scientifica:

> Propongo di considerare la domanda: "Possono le macchine pensare?". Questa indagine dovrebbe iniziare denendo il signicato dei termini "macchina" e "pensare". […] La domanda originaria, "Possono le macchine pensare?", credo sia troppo insensata per meritare di essere discussa. (Turing, 1950)

È un fatto risaputo, anche se talora sottostimato, che le ricerche sull' aspirino **sia a riprodurre i risultati o l'esito positivo** del nostro comportamento intelligente (o almeno di qualche tipo di comportamento animale) con mezzi non biologici, **sia a produrre l'equivalente non biologico della nostra intelligenza**, cioè la fonte di tale comportamento.

Da un lato, come settore dell'ingegneria interessata alla riproduzione del comportamento intelligente, l' ha avuto un successo sbalorditivo.
L' AI riproduttiva ottiene regolarmente risultati migliori e sostituisce l'intelligenza umana in un numero sempre maggiore di contesti.

D'altro lato, come *settore della scienza cognitiva interessata alla produzione di intelligenza*, l'AI rimane fantascienza ed è stata una triste delusione.
L'AI produttiva non si limita a prestazioni inferiori rispetto all'intelligenza umana; non ha ancora preso parte alla competizione.

Oggi, l'AI **scinde la risoluzione ecace dei problemi** e **l'esecuzione corretta** dei compiti dal **comportamento intelligente**, ed è proprio grazie a tale scissione che può incessantemente colonizzare lo spazio sterminato di problemi e compiti, ogni volta che questi possono essere conseguiti senza comprensione, consapevolezza, acume, sensibilità, preoccupazioni, sensazioni, intuizioni, semantica, esperienza, bio-incorporazione, signicato, persino saggezza e ogni altro ingrediente che contribuisca a creare l'intelligenza umana.
In breve, è proprio quando smettiamo di cercare di produrre intelligenza umana che possiamo sostituirla con successo in un numero crescente di compiti.

Se si comprende appieno il senso di questa scissione, si prospettano tre ovvi sviluppi.

1. L'AI dovrebbe smettere di vincere i giochi e imparare a **ludicizzare**. Man mano che l' migliora nel giocare, tutto ciò che può essere trasformato in gioco rientra nel suo ambito.
2. In secondo luogo, in contesti ludicati, l'AI sarà abbinata soltanto all'AI e le sue interazioni interne potrebbero diventare troppo complesse per poter essere integralmente comprese da ammiratori esterni come noi

3. Possiamo aspettarci che l'intelligenza umana abbia un ruolo diverso ovunque l' sia il giocatore migliore. Perché si tratterà meno di risolvere alcuni problemi e più di decidere quali problemi valga la pena di risolvere, perché, per quali nalità, e con quali costi, trade-o e conseguenze accettabili.

In breve, l'AI è definita sulla base di risultati e azioni ingegnerizzati e quindi, nel resto di questo libro, tratterò l'AI come **una riserva di capacità di agire a portata di mano.**

Abbiamo osservato che il digitale sta re-ontologizzando la natura stessa (e quindi il significato) del nostro ambiente, l'infosfera, la quale al contempo sta progressivamente diventando il mondo in cui viviamo.
Quindi, mentre stavamo perseguendo senza successo l'iscrizione dell' produttiva nel mondo, stavamo effettivamente modicando (re-ontologizzando) il mondo per adattarlo all' ingegneristica e riproduttiva.
Il mondo sta diventando un'infosfera sempre meglio adattata alle delimitate capacità dell' AI.

> Per esempio: un robot che dipinge il componente di un veicolo in una fabbrica.
> Lo spazio tridimensionale che definisce i confini entro i quali tale robot può lavorare con successo è denito **l'involucro** del robot.
>
> Alcune delle nostre tecnologie, come le lavastoviglie o le lavatrici, assolvono i loro compiti perché i loro ambienti sono strutturati (**avvolti**) attorno alle capacità elementari del robot al loro interno.
>
> Lo stesso vale, per esempio, per gli scffali robotici nei magazzini di Amazon che sono "avvolti" attorno a loro.

È l'ambiente che è progettato in modo tale da essere compatibile con i robot, non il contrario.

Chi si adatterà a chi?

I sistemi di saranno esponenzialmente più utili ed ecaci nella misura in cui ci inoltreremo nel percorso di digitalizzazione dei nostri ambienti e di espansione dell'infosfera.
L'avvolgimento è una tendenza robusta, cumulativa e che si perfeziona progressivamente.

Il rischio è che potremmo finire per costruire case con pareti rotonde e mobili con gambe abbastanza alte per adattarle alle capacità di Roomba in modo molto più efficace.

Sulla base di questo esempio, è facile percepire come l'opportunità rappresentata dal potere di re-ontologizzazione del digitale si presenti in tre forme: **rifiuto**, **accettazione critica** e **design proattivo**.

Diventando più criticamente consapevoli del potere re-ontologizzante dell'AI e delle applicazioni smart, potremmo essere in grado di evitare le peggiori forme di distorsione (**rifiuto**) o almeno essere coscientemente tolleranti nei loro confronti (**accettazione**), specialmente quando non è importante (penso alla lunghezza delle gambe del divano in casa nostra compatibili con Roomba) o quando si tratta di una soluzione temporanea, in attesa di un design migliore.

In quest’ultimo caso, essere in grado di immaginare come sarà il futuro e quali esigenze di adattamento saranno poste dall’AI e dal digitale più in generale ai loro utenti umani può aiutarci a escogitare soluzioni tecnologiche capaci di diminuire i loro costi antropologici e accrescere i loro benefici ambientali.

In breve, il design umano intelligente (il gioco di parole è voluto) dovrebbe svolgere un ruolo maggiore nel plasmare il futuro delle nostre interazioni con gli artefatti smart attuali e futuri, e gli ambienti che condividiamo con loro.

2.1 L'uso degli esseri umani e delle interfacce

Avvolgere il mondo trasformando un ambiente ostile in un’infosfera adattata digitalmente signica che condivideremo i nostri habitat non solo con forze e fonti di azione naturali, animali e sociali, ma anche e talvolta principalmente con agenti artificiali.

Alcuni dei problemi che stiamo arontando oggi, per esempio, nella sanità digitale o nei mercati nanziari, sorgono già in ambienti altamente [avvolti] in cui tutti i dati rilevanti (e talora gli unici disponibili) sono leggibili da macchine, cosicché decisioni e azioni possono essere compiute automaticamente da applicazioni e attuatori in grado di eseguire comandi e completare le corrispondenti procedure: dall’avvertire o esaminare un paziente all’acquistare o vendere obbligazioni.

Le conseguenze dell’avvolgere il mondo per trasformarlo in un luogo adattato all’ sono molte: un esempio in particolare è molto signicativo e ricco di conseguenze può essere discusso qui

> Gli esseri umani possono diventare inavvertitamente parte del meccanismo.

1. In primo luogo, gli esseri umani stanno diventando nuovi mezzi di produzione digitale
2. Gli esseri umani stanno diventando parte del meccanismo è come clienti inuenzabili
   I servizi “gratuiti” online sono le valute con cui sono “acquistate” le informazioni sui clienti.
   L’AI gioca un ruolo cruciale in questo contesto, ritagliando, ottimizzando e decidendo molti processi attraverso sistemi di raccomandazione

3. Futuro - lo sviluppo prevedibile dell'AI

In precedenza, è stato sostenuto che l'AI non dovrebbe essere interpretata come un matrimonio tra un'intelligenza di tipo biologico e artefatti ingegnerizzati, ma come un *divorzio tra l'agire e l'intelligenza*, cioè una scissione tra la capacità di arontare problemi e compiti con successo in vista di uno scopo e l'esigenza di essere intelligenti nel farlo.

Scrutare nei semi del tempo

Quale futuro possiamo prevedere per l'AI?

Le persone in gamba scommettono su ciò che non è controverso o non può essere vericato. Ciò che è dicile, e potrebbe risultare piuttosto imbarazzante in seguito, è cercare di "scrutare nei semi del tempo, e dire quali chicchi germoglieranno, e quali no", cioè tentare di capire in che direzione è più probabile che l' stia andando o dove potrebbe non andare, dato il suo stato attuale, e su questa base provare a tracciare la mappa delle sde etiche che bisognerebbe prendere sul serio.

Parte della difficoltà è individuare il corretto livello di astrazione, vale a dire identificare l'insieme di osservabili rilevanti ("i semi del tempo") su cui concentrarsi, poiché sono tali osservabili che faranno la vera, signicativa differenza.

Nel nostro caso, sosterrò che i migliori osservabili sono forniti da un'analisi della natura:

a) **dei dati utilizzati** dall'AI per realizzare le proprie prestazioni;
b) **dei problemi che è ragionevole attendersi** che l'IA sia **in grado di risolvere**

Dati storici, ibridi e sintetici e il bisogno di ludicizzazione

Senza dati, gli algoritmi – inclusa l'AI – non vanno da nessuna parte, come un motore con un serbatoio vuoto.
L'AI ha bisogno di dati per essere addestrata e pertanto di dati per applicare il suo addestramento.
Naturalmente, l' può essere estremamente essibile: sono i dati che determinano il suo ambito di applicazione e grado di successo.

È noto che l'AI, intesa come Machine Learning (apprendimento automatico), apprende dai dati che riceve e migliora progressivamente i suoi risultati. Di solito sono necessarie enormi quantità di dati, e di regola quanti più sono i dati, tanto migliore è il risultato.

Tuttavia, recentemente l'AI è talmente migliorata che, in taluni casi, si sta passando da un'enfasi sulla **quantità** di grandi masse di dati, a volte impropriamente chiamati Big Data, a un'enfasi sulla **qualità** di insiemi di dati ben curati.

I "piccoli dati" di alta qualità costituiscono uno degli scenari futuri dell' AI

Ma sappiamo anche che l'AI può generare i propri dati. Chiamerò **sintetici** i dati interamente *generati* dall'AI.

```
> In passato, giocare a scacchi contro un computer signicava giocare contro i migliori giocatori umani che avessero mai preso parte al gioco.
> Ma AlphaZero, l'ultima versione del sistema di sviluppato da DeepMind, ha imparato a giocare meglio di chiunque altro, e in eetti di qualsiasi altro soware, facendo adamento soltanto sulle regole del gioco, senza alcun input di dati da alcuna fonte esterna. Non aveva alcuna memoria storica.
> AlphaZero ha imparato giocando contro se stesso, generando così i propri dati sintetici relativi agli scacchi.
```

AlphaZero ha generato i propri dati sintetici, e questo è stato suffciente per il suo addestramento. Questo è ciò che si intende per dati sintetici.

I dati realmente sintetici hanno alcune straordinarie proprietà, sono:

- durevoli
- riutilizzabili
- rapidamente trasportabili
- facilmente duplicabili
- simultaneamente condivisibili

Sono anche:

- puliti
- affidabili
- non violano privacy o riservatezza
- se vengono persi non è un disastro perché possono essere ricreati
- sono perfettamente formattati per essere utilizzati dal sistema che li genera

Con i dati sintetici l' non è mai costretta ad abbandonare il suo spazio digitale, dove può esercitare il controllo completo su qualsiasi input e output dei suoi processi.

Tra dati storici più o meno mascherati (impoveriti attraverso una risoluzione inferiore, per esempio tramite l'anonimizzazione) e dati puramente sintetici, esiste una varietà di dati più o meno **ibridi**, che possiamo ragurare come un prodotto di dati storici e sintetici.

Un buon esempio, introdotto da Goodfellow e coautori, è fornito dalle reti generative avverse (**GANs**).

Il futuro dell' non risiede soltanto nei "piccoli dati" ma anche, o forse principalmente, nella sua crescente capacità di generare i propri dati.

La differenza è costituita dal processo genetico, cioè dalle regole usate per creare i dati.
I dati storici sono ottenuti tramite *regole di registrazione*, in quanto sono il risultato di osservazioni del comportamento di un sistema.
I dati sintetizzati sono ottenuti tramite *regole di astrazione*, che eliminano, mascherano o ouscano alcuni gradi di risoluzione a partire dai dati storici, per esempio mediante l'anonimizzazione. Dati ibridi e realmente sintetici possono essere generati tramite regole vincolanti o costitutive.

> Quando
> 1. un processo o un'interazione può essere trasformata in un gioco e
> 2. il gioco può essere trasformato in un gioco formato da regole costitutive, allora
> 3. l'AI sarà in grado di generare i propri dati, completamente sintetici, ed essere il miglior "giocatore" su questo pianeta

Problemi difficili, problemi complessi e bisogno di avvolgimento

Allo scopo di comprendere gli sviluppi dell'AI in relazione ad ambienti analogici e digitali, è utile mappare i problemi in base alle *risorse* che sono necessarie per risolverli e capire in che misura l'AI può disporre di tali risorse.
Mi riferisco alle risorse *computazionali* e, pertanto, ai gradi di *complessità*; e alle risorse relative alle *abilità* e, pertanto, ai gradi di *diffcoltà*.

I gradi di **complessità** di un problema sono ben noti e ampiamente studiati nella teoria computazionale.
Conveniamo di mappare la complessità di un problema (trattato dall' AI in termini di spazio-tempo = memoria e passaggi richiesti) da 0 (semplice) a 1 (complesso).

I gradi di **difficoltà** di un problema si basano su una letteratura più qualitativa.
In particolare, ci sono molte maniere per valutare una prestazione e quindi svariati modi per catalogare i problemi relativi alle abilità, ma una distinzione standard è tra abilità motorie **grossolane** e **fini**.
Le abilità grosso-motorie richiedono l'uso di grandi gruppi muscolari; le abilità motorie ni richiedono l'uso di gruppi muscolari più piccoli

Se necessario, utilizzando strumenti della psicologia dello sviluppo, conveniamo di mappare la dicoltà di un problema (trattato dall'AI in termini di abilità richieste) da 0 = facile, a 1 = difficile.

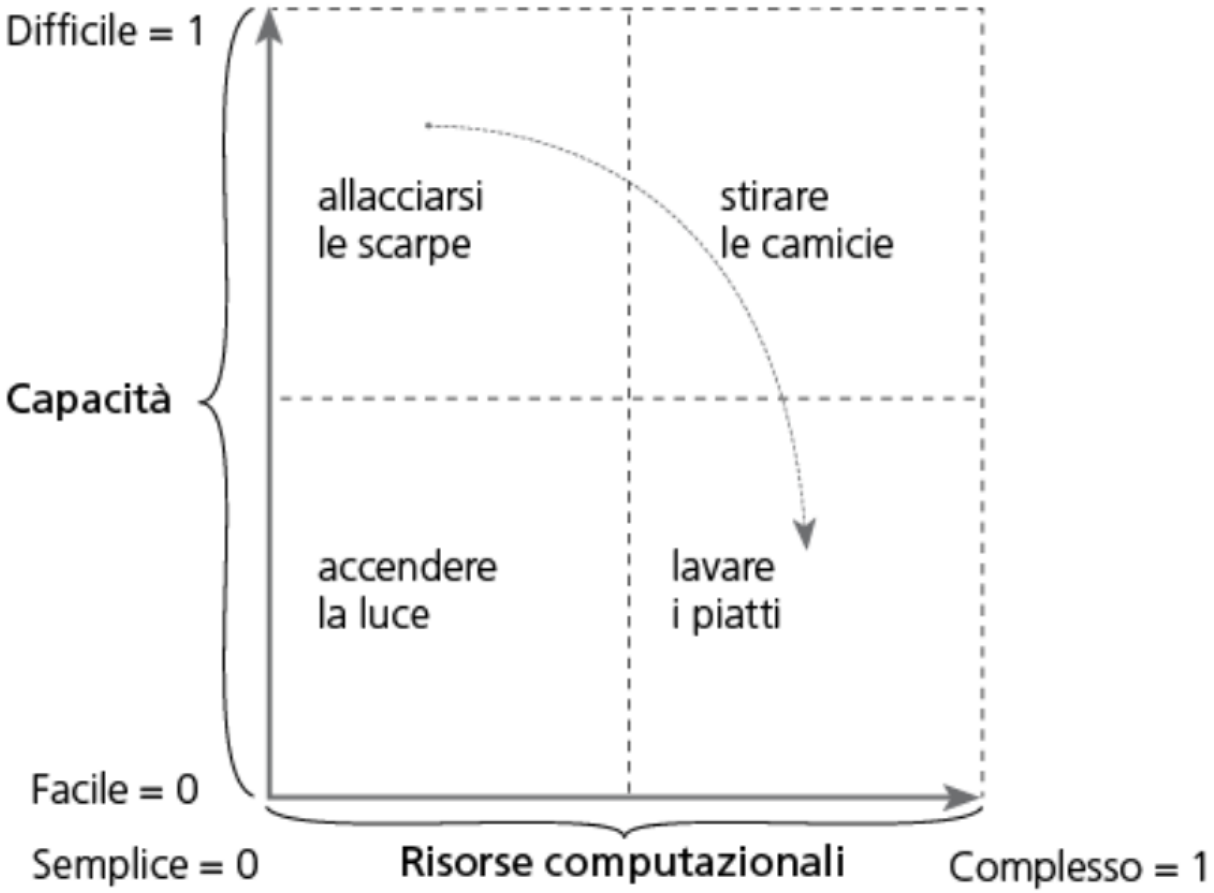


Figura 3.2 Tradurre attività difficili in attività complesse.

> La **diffcoltà** è nemica delle macchine, la **complessità** il loro alleato: per questo, occorre avvolgere il mondo che le circonda, disegnare nuove forme di implementazione per incorporarle con successo nel loro involucro.

3.3 GANs

Due reti neurali – un Generatore e un Discriminatore – competono l’una contro l’altra per avere successo in un gioco.

Lo scopo del gioco per il Generatore è di trarre in inganno il Discriminatore con esempi che paiono simili al set di addestramento.
Quando il Discriminatore rigetta un esempio prodotto dal Generatore, il Generatore impara qualcosa di più su come si presenta un buon esempio.
In altri termini, il Discriminatore fa trapelare informazioni su quanto il Generatore fosse vicino e su come dovrebbe procedere per avvicinarsi.
Col passare del tempo, il Discriminatore impara dal set di addestramento e invia segnali sempre più signicativi al Generatore. Quando ciò si verica, il Generatore si avvicina sempre di più all’apprendimento dell’aspetto degli esempi dal set di addestramento. Ancora una volta, gli unici input che il Generatore ha sono una distribuzione iniziale di probabilità (spesso la distribuzione normale) e l’indicatore che riceve dal Discriminatore. Non vede mai alcun esempio reale.

4. Un quadro unificato di principi etici per l'IA

Troppi principi?

Molte organizzazioni hanno lanciato un'ampia gamma di iniziative per stabilire principi etici per l'adozione di un' socialmente vantaggiosa.

Purtroppo, l'enorme volume di principi proposti rischia di diventare soverchiante e **fuorviante**, sollevando due potenziali problemi.

O i vari insiemi di principi etici per l'AI sono simili, portando a inutili ripetizioni e ridondanze, oppure, se dieriscono in modo signicativo, sono suscettibili di generare confusione e ambiguità.

Un'analisi comparativa di 6 documenti rivela un elevato numero di punti in comune tra gli insiemi di principi esaminati.
Ciò porta a identicare un quadro generale costituito da **cinque principi fondamentali per l'AI etica**.

Ciascun insieme di principi, da cui sono derivati i 5 principi fondamentali, soddisfa quattro criteri di base, per cui è:

a) **recente**, pubblicato a partire dal 2017;
b) direttamente **rilevante** per l'AI e il suo impatto sulla società nel suo insieme
c) di **elevata reputazione**, pubblicato da autorevoli organizzazioni multistakeholder di portata almeno nazionale;
d) **influente**

Questi insiemi vengono confrontati, notando sovrapposizione, con i 4 principi fondamentali della *bioetica*: **beneficenza**, non **maleficenza**, **autonomia** e **giustizia**.
Emerge però l'esigenza di aggiungere un nuovo principio: l'**esplicabilità**.

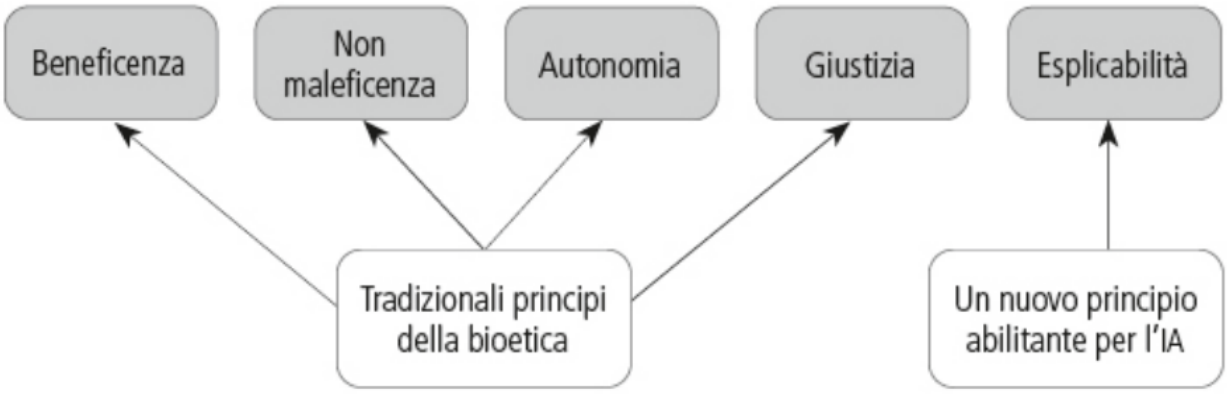


Figura 4.1 Un quadro etico dei cinque principi fondamentali per l'IA.

1. Beneficenza

> "La tecnologia dell' deve essere in linea con l'assicurare le precondizioni di base per la vita sul nostro pianeta, la continua prosperità per l'umanità e la conservazione di un buon

> ambiente per le generazioni future”.

Nel suo insieme, la rilevanza della benecenza sottolinea fermamente l’importanza centrale di promuovere il benessere delle persone e del pianeta con l’AI.

2. Non maleficenza: privacy, sicurezza e cautela della capacità

Benché “Fa’ soltanto del bene” (benecenza) e “Non fare del male” (**non maleficenza**) possano sembrare logicamente equivalenti, non lo sono e rappresentano principi distinti.
I sei documenti incoraggiano tutti la creazione di un’ benefica e ciascuno mette anche in guardia contro le varie conseguenze negative derivanti dall’uso eccessivo o improprio delle tecnologie di IA.

Di particolare interesse è la prevenzione delle violazioni della privacy personale.
Altri mettono in guardia contro le minacce di una corsa agli armamenti di AI e dell’automiglioramento ricorsivo dell’AI.

3. Autonomia

Quando adottiamo l’AI e il suo agire smart, cediamo volontariamente parte del nostro potere decisionale ad artefatti tecnologici.

> Per questo, affermare il principio di **autonomia** nel contesto dell’AI signica trovare un equilibrio tra il potere decisionale che ci riserviamo e quello che deleghiamo agli agenti articiali. Il rischio è che la crescita dell’**autonomia artificiale** possa minare il orire dell’**autonomia umana**.

È chiaro dunque sia che l’autonomia umana debba essere promossa, sia che l’autonomia delle macchine debba essere limitata e resa intrinsecamente reversibile, qualora l’autonomia umana debba essere protetta o ristabilita.

> Ciò introduce una nozione che può essere definita come **meta-autonomia**, o *modello di decisione di delega*.
> Gli esseri umani dovrebbero mantenere il potere di decidere quali decisioni prendere, e quali delegare, ma qualsiasi delega dovrebbe rimanere rivedibile, adottando come ultima garanzia **il potere di decidere di decidere di nuovo**.

4. Giustizia: promuovere la prosperità, preservare la solidarietà, evitare l'iniquità

> "Lo sviluppo dell’AI dovrebbe promuovere la giustizia e cercare di eliminare tutti i tipi di discriminazione"

> L'IA dovrebbe "contribuire alla giustizia globale e alla parità nell’accesso ai benefici"

> Altrove “giustizia” ha ancora altri signicati (soprattutto nel senso di **equità**), variamente collegati all’uso dell’ per correggere errori del passato come eliminare discriminazioni ingiuste, promuovere la diversità e prevenire l’insorgenza di nuove minacce alla giustizia.

5. Esplicabilità: rendere possibili gli altri principi tramite l'intelligibilità e la responsabilità

Il funzionamento dell'AI è spesso invisibile o incomprensibile a tutti tranne (nella migliore delle ipotesi) agli osservatori più esperti.

Per questo, tutti fanno riferimento alla

> necessità di comprendere e di rendere conto dei processi decisionali dell' AI.

L'aggiunta del principio di "esplicabilità", che include sia il senso epistemologico di "intelligibilità" sia il senso etico di "responsabilità", è il cruciale pezzo mancante del puzzle etico dell'AI.

5. Dai principi alle pratiche - i rischi di comportamenti contrari all'etica

È tempo che il dibattito sull'etica dell'AI evolva dal *cosa* al *come*: non si tratta soltanto di individuare quale etica sia necessaria, ma anche come l'etica possa essere applicata e implementata in modo effcace e con successo per cambiare le cose in meglio.

Tuttavia, nel tradurre i principi etici in buone pratiche, anche i migliori sforzi possono essere minati da alcuni rischi di comportamenti contrari all'etica.

1. Lo Shopping Etico

Un rischio legato all'etica è che l'iperattività nella definizione dei principi crei un "mercato di principi e valori" in cui attori pubblici e privati possano acquistare il tipo di etica che meglio si adatta a giusticare i loro comportamenti attuali, piuttosto che rivedere questi comportamenti per renderli coerenti con un quadro etico socialmente condiviso.

> *Shopping etico digitale* = def. il malcostume di scegliere, adattare o rivedere ("mescolare e abbinare") principi etici, linee guida, codici, quadri di riferimento o altri standard simili (specialmente ma non solo nell'etica dell'), estraendoli da una varietà di offerte disponibili, per conferire una patina nuova ad alcuni comportamenti preesistenti (scelte, processi, strategie ecc.) e in tal modo giustificarli a posteriori, invece di implementare o affinare nuovi comportamenti confrontandoli con standard etici pubblici.

La strategia per affrontare lo shopping etico digitale è quella di stabilire standard etici chiari, condivisi e pubblicamente accettati.

2. Il bluewashing etico

Nell'etica ambientale, il **greenwashing** è il malcostume di un attore pubblico o privato che cerca di apparire più verde, più sostenibile o più rispettoso dell'ambiente di quanto non sia in realtà.

> *Bluewashing etico* = def. il malcostume di fare affermazioni infondate o fuorvianti al riguardo (o di attuare misure superficiali a favore) dei valori etici e dei benefici di processi, prodotti, servizi o altre soluzioni digitali al fine di apparire più etici dal punto di vista digitale di quanto non si sia effettivamente.

Il greenwashing e il bluewashing etici sono entrambi forme di disinformazione, spesso ottenute impiegando una frazione delle risorse che sarebbero necessarie per arontare i problemi etici che pretendono di affrontare.

La migliore strategia contro il bluewashing è la stessa già adottata contro il greenwashing: trasparenza e formazione

3. Il lobbismo etico

Gli attori privati cercano, talvolta, di utilizzare l'autoregolazione (o almeno si sospetta che lo facciano) nell'ambito dell'etica dell' per esercitare pressioni (azioni di **lobbying**) contro l'introduzione di norme giuridiche, oppure a favore del loro annacquamento, di una loro applicazione più attenuata, o inne per fornire una giusticazione a una loro limitata osservanza

> *Lobbismo etico digitale* = def. il malcostume di sfruttare l'etica digitale per ritardare, rivedere, sostituire o evitare un'idonea e necessaria regolazione giuridica (o la sua applicazione) relativa al design, lo sviluppo e l'implementazione di processi, prodotti, servizi o altre soluzioni digitali.

La strategia contro il lobbismo etico digitale è duplice:

- **deve essere contrastato da una buona normativa** e da una sua applicazione efficace.
- il lobbismo etico digitale **deve essere reso manifesto** ogni volta che si verica ed essere chiaramente distinto dalle vere forme di autoregolazione.

4. Il dumping etico

"**Dumping etico**" è un'espressione coniata nel 2013 dalla Commissione europea per descrivere l'*esportazione di pratiche di ricerca contrarie all'etica di paesi dove vi siano cornici legali ed etiche e meccanismi di applicazione delle norme più deboli.*

In ambiti digitali lo spettro di regimi giuridici e cornici etiche facilita l'esportazione di pratiche contrarie all'etica (o addirittura illegali) e l'importazione dei risultati di tali pratiche.
In altre parole, il problema è duplice, di **etica della ricerca** e di **etica del consumo**.

> *Dumping etico digitale* = def. il malcostume di
> (a) esportare attività di ricerca riguardo a processi, prodotti, servizi o altre soluzioni digitali, in altri contesti o luoghi (per esempio, da organizzazioni europee al di fuori della UE) in modi che sarebbero eticamente inaccettabili nel contesto o luogo di origine; e di
> (b) importare i risultati di tali attività di ricerca contrarie all'etica

Anche in questo caso la strategia è duplice. Bisogna concentrarsi sull'etica della ricerca e sull'etica del consumo. Se si vuole essere coerenti, entrambe devono ricevere pari attenzione.

5. L'elusione dell'etica

> Elusione dell'etica = def. il malcostume di svolgere sempre meno "lavoro etico" (come adempiere ai doveri, rispettare i diritti, onorare gli impegni ecc.) in un dato contesto quanto più basso è percepito (erroneamente) il ritorno di tale lavoro etico in quel contesto.

La strategia contro l'elusione dell'etica consiste nell'arontare la sua origine, che è la mancanza di una chiara allocazione di responsabilità. Gli agenti possono essere tanto più tentati di sottrarsi al loro impegno etico in un dato contesto, quanto più reputano di poter trasferire le responsabilità altrove.

Ciò accade più frequentemente e facilmente nei “contesti D”, dove la propria responsabilità viene percepita (erroneamente) come meno elevata perché *distante, diminuita, delegata o distribuit*a.

6. Etica soft e governance dell'AI

Che tipo di mature società dell'informazione vogliamo costruire? Qual è il nostro progetto umano per l'era digitale?

Guardando indietro al nostro presente, questo è il momento storico in cui si vedrà che abbiamo gettato le basi per le nostre mature società dell'informazione.
Saremo giudicati dalla qualità del nostro lavoro.
Per questo, chiaramente, la vera sfida non è la buona innovazione digitale, ma la buona governance del digitale.

Etica, regolazione e governance

Sulla governance delle tecnologie digitali in generale e dell' AI in particolare un punto è chiaro:

i) **la governance digitale**
ii) **l'etica digitale**
iii) **la regolazione digitale**

sono approcci normativi *diversi, complementari*, da non confondere tra loro, ma da tenere chiaramente distinti.

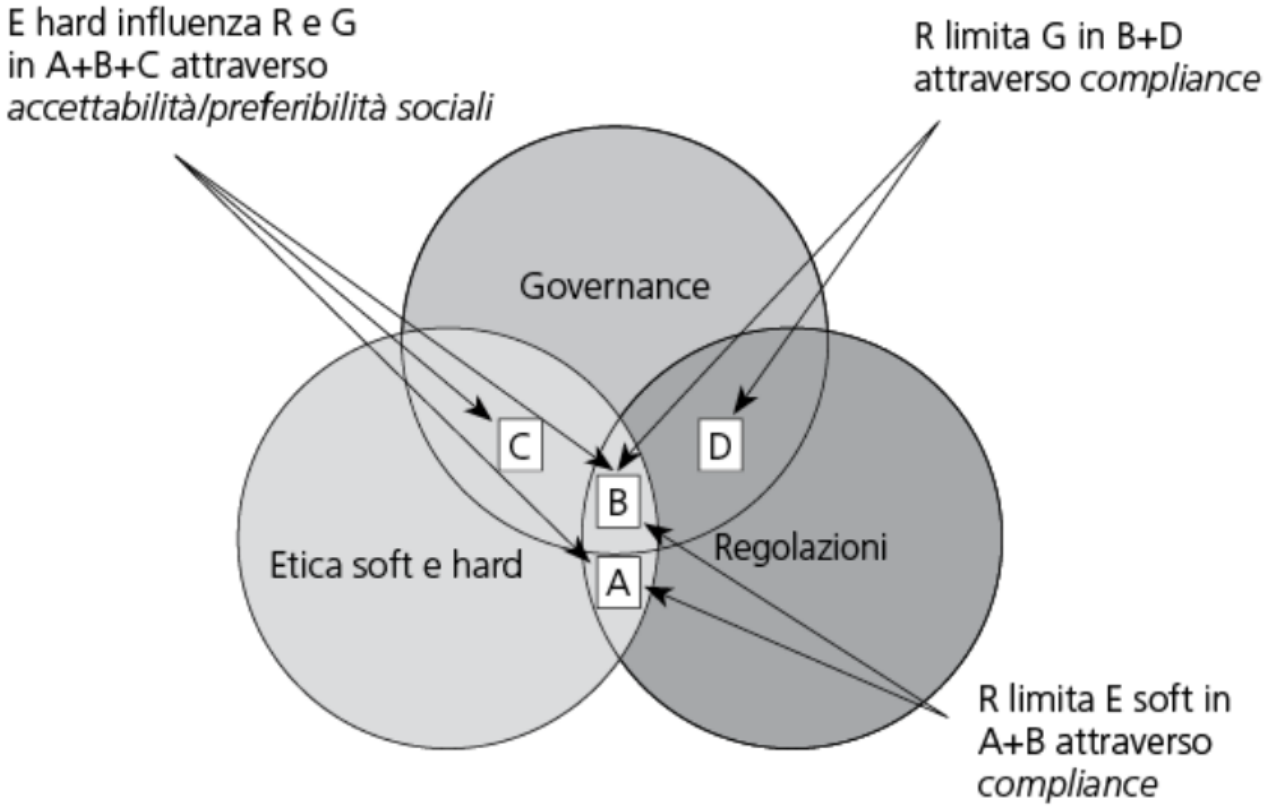


Figura 6.1 Etica digitale, regolazioni digitali e governance digitale.

> La **governance digitale** è la pratica di stabilire e attuare politiche, procedure e standard per i corretti sviluppo, utilizzo e gestione dell'infosfera.

La governance digitale può comprendere linee guida e raccomandazioni che si sovrappongono alla **regolazione digitale** ma non sono identiche a essa.

> La **regolazione digitale** è solo un altro modo di riferirsi alla legislazione pertinente, un sistema di leggi elaborato e applicato attraverso istituzioni sociali o governative per regolare il comportamento degli agenti rilevanti nell'infosfera.

Non tutti gli aspetti della regolazione digitale sono una questione di governance digitale e non tutti gli aspetti della governance digitale sono una questione di regolazione digitale

La **compliance** (vale a dire la conformità alle norme) è la relazione cruciale attraverso la quale la regolazione digitale modella la governance digitale.

> Tutto ciò vale per l'**etica digitale**, intesa come quel settore dell'etica che studia e valuta i problemi morali relativi a
>
> - dati e informazioni
> - (inclusi generazione, registrazione, cura, trattamento, diusione, condivisione e utilizzo)
> - algoritmi
> - (tra cui AI, agenti articiali, e robot)
> - le relative pratiche e infrastrutture
> - (inclusi innovazione responsabile, programmazione, hackeraggio, codici professionali e standard)
>
> al fine di formulare e supportare soluzioni moralmente buone

L'**etica digitale** modella la regolazione digitale e la governance digitale attraverso la relazione di valutazione morale di ciò che è socialmente accettabile o preferibile.

Quando i decisori politici si chiedono perché dovremmo impegnarci in valutazioni etiche quando la compliance è già presupposta, la risposta dovrebbe essere chiara: la compliance è necessaria ma insufficiente per guidare la società nella giusta direzione.

> Perché la **regolazione digitale** indica quali sono le mosse **valide** e non valide nel gioco, ma non dice nulla su quali potrebbero essere le mosse **buone** o **migliori**, tra quelle valide per **avere una società migliore**.
>
> Questo è il compito sia dell'etica digitale, sul lato dei valori e delle preferenze morali, sia della buona governance digitale, sul lato della gestione

Etica HARD e SOFT

L'etica digitale può dunque essere intesa in due modi, come **etica hard** o **soft**.

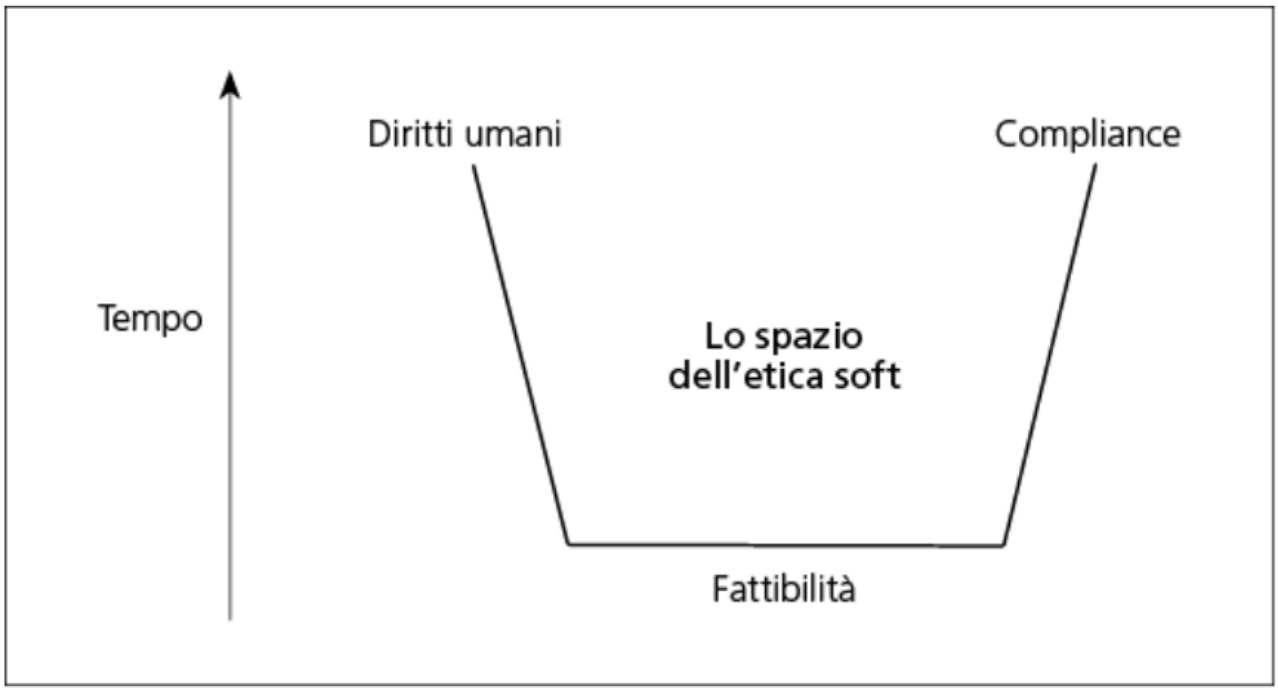


Figura 6.2 Lo spazio dell'etica soft.

> L'**etica hard** è ciò che di solito abbiamo in mente quando discutiamo di valori, diritti, doveri e responsabilità – o, più in generale, di ciò che è moralmente giusto o sbagliato, di ciò che dovrebbe o non dovrebbe essere fatto – quando formuliamo nuove normative o sottoponiamo a critica quelle esistenti.
> ***In breve, nella misura in cui l'etica contribuisce a creare, plasmare o modicare il diritto, possiamo chiamarla etica hard.***

> L'**etica soft** comprende lo stesso ambito normativo dell'etica hard, ma lo fa considerando ciò che dovrebbe o non dovrebbe essere fatto **al di là** della normativa vigente, non contro di essa, o nonostante il suo ambito di applicazione, o per cambiarla.
>
> Pertanto, l'etica so può includere *l'autoregolazione*.
>
> In altre parole, ***l'etica soft è un'etica post-compliance perché, in questo caso, "il dover fare qualcosa implica il poter fare quel qualcosa".***

L'etica soft come quadro etico

È giunto il momento di fornire un'analisi più specica, e per questo si farà affidamento sul GDPR. La scelta sembra ragionevole: dato che la regolazione digitale nella è ora determinata dal , e che la normativa della è solitamente rispettosa dei diritti umani, può essere utile comprendere il valore della distinzione tra etica so e hard e le loro relazioni con il diritto utilizzando il come caso concreto di applicazione.

Per comprendere il ruolo dell'etica hard e so rispetto al diritto in generale e al in particolare, è necessario introdurre cinque elementi.

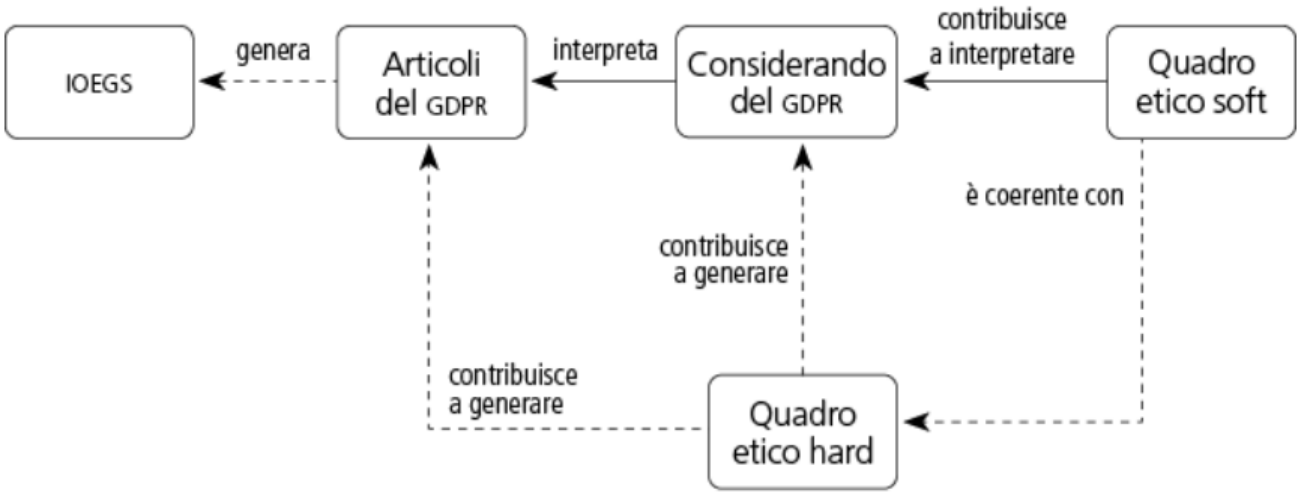


Figura 6.3 Etica soft e hard e la loro relazione con la regolamentazione. Si noti che il diagramma viene semplificato omettendo i riferimenti a tutti gli altri elementi che contribuiscono ai vari quadri di riferimento.

- Implicazioni etiche, giuridiche e sociali (IEGS) del GDPR
- Articoli del GDPR stesso
  - progettata per
    - armonizzare le norme sulla protezione dei dati personali in tutta Europa
    - proteggere e far rispettare la privacy dei dati di tutti i cittadini della
    - per migliorare il modo in cui le organizzazioni in tutta la UE affrontano la privacy dei dati
  - comprende 99 articoli
    - gli articoli non comprendono tutto, lasciano zone grigie di incertezza normativa
      - sono soggetti a interpretazioni e possono richiedere un aggiornamento
        - per questo sono accompagnati dai **considerando**
- 173 **considerando**:
  - testi che spiegano le ragioni delle disposizioni di un atto; **non sono** giuridicamente vincolanti e non dovrebbero contenere un linguaggio normativo
  - sono utilizzati dalla Corte di Giustizia dell'Unione Europea per interpretare una direttiva o un regolamento e adottare una decisione nel contesto di un caso concreto.
  - anche i considerando richiedono un'interpretazione
    - **quadro etico**
- **quadro etico soft**
  - contribuisce, insieme ad altri fattori, alla comprensione dei considerando
- **quadro etico hard**
  - l'elemento etico (insieme ad altri) che ha motivato e guidato il processo che ha portato all'elaborazione della legge

Chiaramente, il ruolo dell'etica è sia precedente sia successivo alla legge, in quanto contribuisce prima a renderla possibile e in seguito a integrarla (costringendola talora anche a cambiare).

> In tale contesto, si può sostenere che il diritto *contiene non solo regole ma anche principi*.
>
> Soprattutto nei casi diffcili, poco chiari o non disciplinati, in cui le regole non riescono a essere applicabili in modo completo o inequivocabile a una situazione concreta la decisione del caso è e dovrebbe essere guidata da principi di etica soft.

Analisi dell'impatto etico

> Dato il futuro aperto dell'etica digitale, è ovvio che l'analisi dell'impatto (AIE) etico debba diventare una priorità.

Oggi, l'AIE può essere basata sull'analisi dei dati applicata strategicamente alla valutazione dell'impatto etico di tecnologie, beni, servizi e pratiche digitali.

È cruciale perché il compito dell'etica digitale non è soltanto quello di "scrutare nei semi digitali del tempo / e dire quali chicchi germoglieranno, e quali no", ma anche quello di cercare di determinare quali dovrebbero crescere e quali no.

Figura 6.4 Analisi dell'impatto etico (AIE): il ciclo di analisi di previsione.

> **Dobbiamo anticipare e guidare lo sviluppo etico dell'innovazione tecnologica.**
>
> Lo possiamo fare
> - valutando ciò che è effettivamente fattibile
> - privilegiando, al suo interno, ciò che è sostenibile dal punto di vista ambientale
> - uindi ciò che è socialmente accettabile e
> - infine, idealmente, scegliendo ciò che è socialmente preferibile

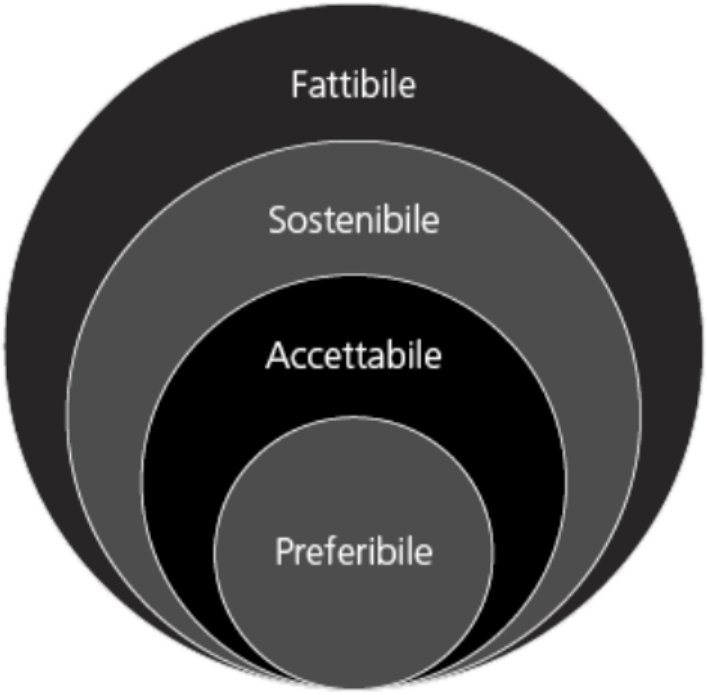


Figura 6.5 Valutazione dell'impatto dell'etica digitale.

Preferibilità digitale e cascata normativa

*Non disponiamo ancora, per l'infosfera, di un concetto equivalente alla **sostenibilità** per la **biosfera**, perciò la nostra attuale equazione è incompleta*

Un potenziale candidato potrebbe essere "**equità**".

Eppure la mancanza di terminologia concettuale non rende la buona governance del digitale meno urgente o un mero sforzo utopico. In particolare, l'etica digitale, con i suoi valori, principi, scelte, raccomandazioni e vincoli, inuenza già in modo signicativo, e talvolta molto più di ogni altra forza, il mondo della tecnologia.

> Sul lungo termine, le persone sono vincolate in ciò che possono o non possono fare (**possibilità**) -> dai beni e servizi forniti dalle organizzazioni che sono vincolate -> dal diritto (**compliance**), ma quest'ultimo è modellato e vincolato -> dall'etica

Purtroppo, una tale **cascata normativa** diventa palese soprattutto quando si verica un contraccolpo, cioè specialmente in contesti negativi, quando il pubblico riuta alcune soluzioni, anche laddove possono essere buone soluzioni.

Una cascata normativa dovrebbe invece essere utilizzata in modo costruttivo, per perseguire la costruzione di una società dell'informazione matura di cui essere orgogliosi.

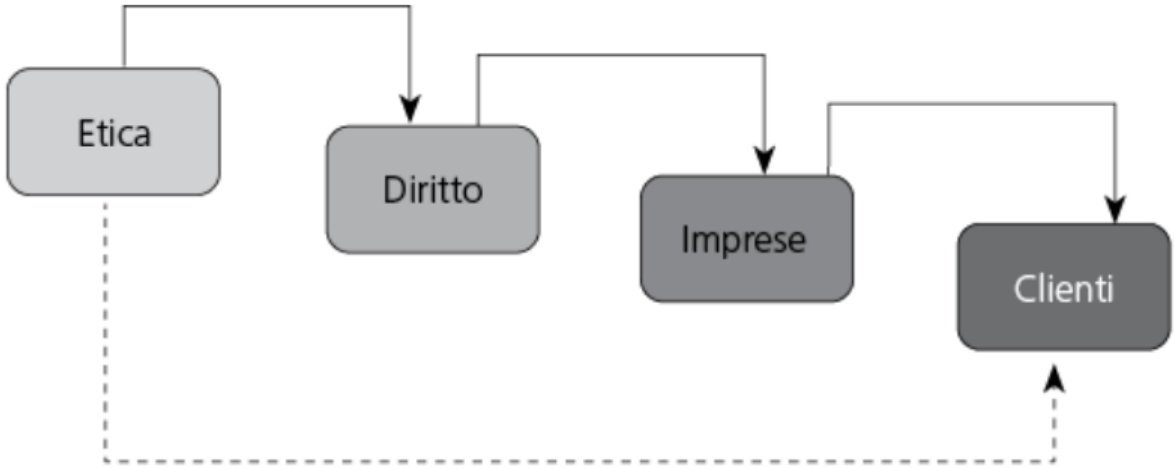


Figura 6.7 Esempio di cascata normativa. L'esempio usa le imprese come agenti e le persone come clienti. Le imprese potrebbero essere sostituite dal governo e le persone dai cittadini.

> Il rispetto delle norme è senz'altro necessario, ma largamente insuffciente. L'adozione di un approccio **etico** all'innovazione digitale conferisce quello che può essere denito un "duplice vantaggio"

- Da un lato, l'etica soft può fornire una strategia di opportunità, consentendo agli attori di sfruttare il valore sociale delle tecnologie digitali
- D'altra parte, l'etica fornisce anche una **soluzione per la gestione del rischio**, in quanto consente alle organizzazioni di anticipare ed **evitare errori costosi**

In tal modo, l'etica può anche abbassare i costi derivanti dall'opportunità delle scelte non compiute o delle opzioni non colte per paura di sbagliare.

Il duplice vantaggio dell'etica so può funzionare solo in un contesto di legislazione adeguata, fiducia pubblica e responsabilità chiare in senso più ampio**

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.

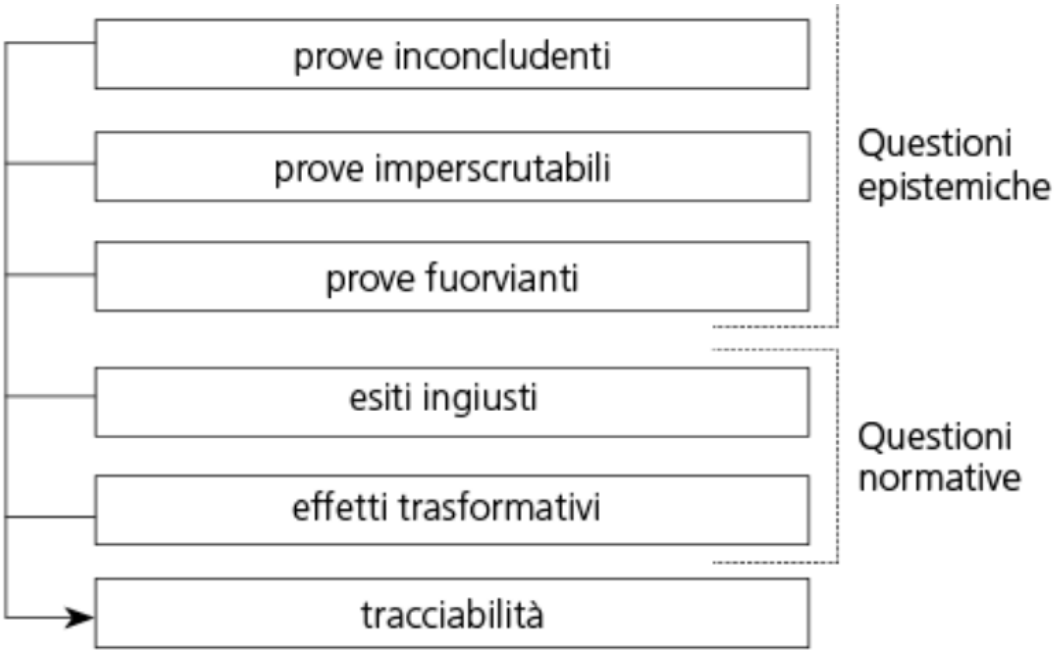


Figura 7.1 Sei tipi di questioni etiche sollevate dagli algoritmi (Mittelstadt, Allo, Taddeo et al., 2016, p. 4).

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.

7.1 Privacy Differenziale

> La privacy differenziale consente ai ricercatori e agli analisti di database di ottenere informazioni preziose dai database senza divulgare le informazioni di identificazione personale degli individui. Questo è fondamentale poiché molti database contengono una varietà di informazioni personali.

Il modo in cui funziona la privacy differenziale consiste nell'introdurre una **perdita di privacy o un parametro di budget per la privacy**, spesso indicato come epsilon (ε), nel set di dati. Questi parametri controllano quanto rumore o casualità viene aggiunto al set di dati non elaborato.

Ad esempio, si immagini di avere una colonna nel set di dati con le risposte "Sì"/"No" delle persone.

Si supponga ora di lanciare una moneta per ogni individuo:

- **Teste:** la risposta è lasciata così com'è.
- **Code:** si capovolge una seconda volta, registrando la risposta come "Sì" se testa e "No" se croce, indipendentemente dalla risposta reale.

Utilizzando questo processo, aggiungi casualità ai dati. Con una grande quantità di dati e le informazioni dal meccanismo di aggiunta del rumore, il set di dati rimarrà accurato in termini di misurazioni aggregate. La privacy entra in gioco consentendo a ogni singolo individuo di negare plausibilmente la propria vera risposta grazie al processo di randomizzazione.

8. Cattive pratiche - l'uso improprio dell'AI per il male sociale

L'uso criminale dell'IA

L' IA può svolgere un ruolo sempre più **essenziale** negli atti criminali in futuro.
Esempi chiari di che vengono chiamati “crimini di AI” (**CIA**) sono forniti da due esperimenti di ricerca (teorici).

Nel primo, due scienziati sociali computazionali hanno usato l’AI come strumento per convincere utenti di social media a cliccare su collegamenti di phishing; poiché ogni messaggio è stato costruito con tecniche di ML applicate ai comportamenti passati e ai profili pubblici degli utenti, il contenuto è stato ritagliato su ciascun individuo.

Nel secondo esperimento, tre scienziati informatici hanno simulato un mercato e hanno scoperto che gli agenti di scambio potevano apprendere ed eseguire una “vantaggiosa” campagna di manipolazione del mercato che includeva una serie di falsi ordini ingannevoli.

Di seguito, l'analisi riportata risponde a due domande:

1. *Quali sono le minacce fondamentalmente peculiari e plausibili poste dai CIA?*
2. *Quali soluzioni sono disponibili o possono essere elaborate per affrontare i CIA?*

Preoccupazioni

Un'analisi iniziale di revisione della letteratura ha filtrato i risultati relativi ad atti o omissioni criminali che:

- si sono vericati o probabilmente si vericheranno in base alle attuali tecnologie di (**plausibilità**);
- richiedono l’ come fattore essenziale (**unicità**);
- sono perseguiti nel diritto nazionale

Tabella 8.1 Mappa delle minacce settoriali e trasversali, basata sulla revisione della letteratura.

Aree criminali	**Motivi di preoccupazione**			
	Emergenza	**Responsabilità**	**Monitoraggio**	**Psicologia**
Commercio, mercati finanziari e insolvenza	✓	✓	✓	
Droghe nocive o pericolose			✓	✓
Reati contro la persona	✓	✓		
Reati sessuali				✓
Furto e frode, contraffazione e sostituzione di persona			✓	

Ciò ha portato a individuare cinque aree criminali potenzialmente interessate dai CIA:

1. **commercio, mercati finanziari e insolvenza**
2. **droghe** nocive o pericolose
3. reati **contro la persona** (inclusi omicidio doloso o colposo, molestie, stalking, tortura);
4. reati **sessuali** (compresi stupro, aggressione sessuale);
5. **furto e frode**, contraffazione e sostituzione di persona.

Vengono identificate 4 ragioni di preoccupazione:

1. **Emergenza**

> ```
> La preoccupazione per l'emergenza si riferisce al fatto che un'analisi superciale del design e dell'implementazione di un agente articiale potrebbe suggerire un tipo particolare di comportamento relativamente semplice, ma la verità è che, a seguito dell'implementazione, l'AI può agire in modi potenzialmente più sosticati che vanno oltre le nostre aspettative iniziali.
> Pertanto, azioni e piani coordinati possono **emergere autonomamente**.
>
> Il comportamento emergente potrebbe avere implicazioni criminali, nella misura in cui devia dal design originale
> ```

2. **Responsabilità**

> ```
> La preoccupazione relativa alla responsabilità si riferisce al fatto che i CIA potrebbero minare i modelli di responsabilità esistenti, minacciando così il potere dissuasivo e riparatore della legge.
>
> La prima condizione per la responsabilità penale è l'**actus reus**: un atto o un'omissione criminale posta in essere **volontariamente**.
> Per le tipologie di deniti in modo tale che solo l' può realizzare l'atto o l'omissione criminale, l'aspetto volontario dell'actus reus potrebbe non essere mai soddisfatto poiché l'idea che un possa agire volontariamente è priva di fondamento.
>
> Quando la responsabilità penale è basata sulla colpa, ha anche una seconda condizione, la **mens rea** (una mente colpevole). ^b17b93 ^bdc55f ^8009a2 ^c9f83d
> ```

Un agente articiale può essere **responsabile causalmente** di un atto criminale, ma soltanto un agente umano può esserne **moralmente responsabile**.

> La complessità dell’AI fornisce un grande incentivo agli agenti umani per evitare di scoprire cosa sta facendo esattamente il sistema di AI, poiché meno gli agenti umani sanno, più saranno in grado di negare la loro responsabilità

In alternativa, i legislatori possono denfiire la responsabilità penale **senza un requisito di colpa**; ciò porterebbe ad attribuire la responsabilità alla persona giuridica senza colpa che ha attivato un nonostante il rischio che possa plausibilmente compiere un’azione o un’omissione criminale.
La responsabilità si applica agli agenti che **fanno la differenza** in un sistema complesso in cui i singoli agenti svolgono azioni neutrali che però sfociano in un crimine collettivo.

3. **Monitoraggio**

> ```
> La preoccupazione per il monitoraggio dei fa riferimento a tre tipi di problemi: **attribuzione, fattibilità e azioni intersistemiche**.
>
> > L’**attribuzione** del mancato rispetto della normativa vigente costituisce un problema di monitoraggio degli utilizzati come strumenti di reato, dovuta alla capacità di questa nuova tipologia di agenti smart di operare in modo indipendente e autonomo: due caratteristiche che tendono a confondere ogni tentativo di tracciare la responsabilità riconducendo gli effetti di un’azione all’autore del reato. ^7b4251
> ```

> ```
> > Per quanto riguarda la fattibilità del monitoraggio, l’autore di un reato può trarre vantaggio dai casi in cui gli operano a velocità e livelli di complessità che vanno semplicemente al di là della capacità di monitorarne la conformità con le norme.
>
> > Le azioni intersistemiche fanno riferimento a un problema per i sistemi di monitoraggio dei con visione a tunnel che si **concentrano solo su un singolo sistema**.
> ```

4. **Psicologia**

> ```
> La **psicologia** fa riferimento alla preoccupazione che l’ possa inuenzare/manipolare negativamente lo stato mentale di un utente no al punto di agevolare o causare (in tutto o in parte) il crimine.
> Un effetto psicologico si basa sulla capacità degli di ottenere la ducia degli utenti, rendendo le persone vulnerabili alla manipolazione.
> ```

Minacce

1. Commercio, mercati finanziari e insolvenza

> Attualmente, sorgono problemi nel caso del coinvolgimento dell' soprattutto in tre aree: **manipolazione del mercato, fissazione dei prezzi e collusione**.

La **manipolazione del mercato** è definita come quelle "azioni e/o operazioni da parte di partecipanti al mercato che tentano di infuenzare artificialmente i prezzi di mercato".

È stato dimostrato che tali forme di inganno emergono da un'implementazione apparentemente conforme di un progettato AA per operare per conto di un utente (cioè un agente articiale di trading).
Questo perché un AA, in particolare uno che apprende da osservazioni reali o simulate, può imparare a generare segnali che sono effettivamente ingannevoli:

- effettuare ordini senza alcuna intenzione di eseguirli, semplicemente per manipolare gli onesti partecipanti al mercato
- acquisire una posizione in uno strumento finanziario, come un titolo, per poi gonare articialmente il titolo tramite la sua promozione fraudolenta

Questo è noto in termini colloquiali come schema "pompa e sgona" (**pump-and-dump**).

La **collusione**, sotto forma di **fissazione dei prezzi**, può emergere anche nei sistemi automatizzati grazie alle capacità di pianificazione e autonomia degli AA:

- algoritmi imparano a coordinarsi, ad esempio per tenere un prezzo alto

L'assenza di intenzionalità, l'intervallo decisionale molto breve e la probabilità che la collusione emerga a seguito delle interazioni tra sollevano anche serie preoccupazioni per quanto riguarda la responsabilità e il monitoraggio.

2. Droghe nocive o pericolose

> I crimini che rientrano in questa categoria includono il **traffico, la vendita, l'acquisto e il possesso di droghe vietate.**

In questo caso, l'AI può fungere da strumento per sostenere il traco e la vendita di sostanze illecite.
Il traffco business-to-business di droga che utilizza l'AI è una minaccia dovuta ai criminali che adoperano **veicoli senza equipaggio**, che fanno leva sulla pianicazione dell'AI e sulle tecnologie di navigazione autonoma come strumenti per migliorare i tassi di successo del contrabbando.

Poiché le reti di contrabbando vengono fermate dal monitoraggio e dall'intercettazione delle linee di trasporto, l'applicazione delle norme diventa più dicile quando vengono usati veicoli senza equipaggio per trasportare ciò che è contrabbandato.

3. Reati contro la persona

> I crimini che rientrano nella categoria dei reati contro la persona riguardano **molestie** e **torture**.

> Le **molestie** comprendono comportamenti intenzionali e ripetitivi che generano allarme o causano disagio a una persona.

Per quanto riguarda i CIA basati sulle molestie, la letteratura fa riferimento ai social bot. Un malintenzionato può avvalersi di un social bot come strumento di molestia diretta o indiretta. La molestia diretta è costituita dalla diffusione di messaggi di odio contro la persona.

> Per quanto riguarda la tortura, si configura quando un pubblico ufficiale infligge intenzionalmente gravi dolori o sofferenze a un altro soggetto nell'esercizio o nel presunto esercizio delle sue funzioni ufficiali.

L'uso dell'IA per l'interrogatorio è motivato dalla sua capacità di rilevare meglio l'inganno, l'emulazione dei tratti umani (come la voce) e la modellazione affettiva per manipolare l'interrogato.
Tuttavia, un con queste capacità può imparare a **torturare una vittima**.

L'interrogato probabilmente sa che l'AI non può comprendere il dolore o provare empatia, ed è pertanto improbabile che agisca con pietà e interrompa l'interrogatorio.
Senza compassione la semplice presenza di un di interrogatorio può far capitolare il soggetto per paura, il che, secondo il diritto internazionale, potrebbe costituire un **crimine di tortura** (minacciata).
Inoltre, chi si avvale di un AA può essere in grado di distaccarsi, **emotivamente e fisicamente**; perciò, diventa più facile ricorrere alla tortura.

4. Reati sessuali

> I reati sessuali discussi in letteratura in relazione all' sono i seguenti: stupro (cioè sesso penetrativo senza consenso), aggressione sessuale (cioè contatto sessuale senza consenso) e rapporti o attività sessuali con un minore.

Questi crimini coinvolgono l'AI quando, tramite un'interazione avanzata uomo-computer, quest'ultima **promuove l'oggettivazione sessuale o l'abuso e la violenza sessualizzati**, e potenzialmente simula e quindi **aumenta il desiderio sessuale** per i reati sessuali.

5. Furto e frode, contraazione e sostituzione di persona

> Contraffazione e sostituzione di persona sono collegate tramite i CIA a furti e frodi extra-aziendali, con implicazioni anche per l'uso di nelle frodi aziendali.

Per quanto riguarda il furto e la frode extra-aziendale, il processo prevede due fasi.

- Inizia con l'utilizzo dell'AI per raccogliere dati personali
  - usando social bot di social media
  - phishing
- procede con l'utilizzo dei dati personali rubati e di altri metodi di AI per forgiare un'identità che induca le autorità bancarie a eettuare una transazione (*ovvero furto e frode bancaria*)

Soluzioni disponibili

1. Affrontare l'emergenza

> Le soluzioni giuridiche possono comportare la limitazione dell’autonomia degli agenti o del loro impiego.
> - Per esempio, la Germania ha creato contesti deregolamentati in cui è consentita la sperimentazione di automobili a guida autonoma

2. Affrontare la responsabilità

4 modelli:

- **responsabilità diretta**: attribuisce gli elementi fattuali e mentali a un AA
  - un limite fondamentale di questo modello risiede nel fatto che gli AA non hanno personalità giuridica e capacità di agire, e quindi non possiamo ritenere un AA legalmente responsabile
  - porterebbe a una deresponsabilizzazione degli agenti umani dietro l'AA
- **perpetrazione per mezzo di altri**: l'AA è uno strumento il cui orchestratore è il vero autore
  - 3 candidati umani: programmatori, produttori e utilizzatori
    - per essere responsabile, l’operatore di un deve volere la realizzazione del fatto illecito
- **responsabilità di comando**: nei contesti in cui esiste una catena di comando, attribuisce la responsabilità a **qualsiasi ufficiale che sia a conoscenza ma non si adopera per prevenire i crimini**
  - Tuttavia questioni relative a livelli di crescente complessità nella programmazione, relazioni robo-umane e integrazione in strutture gerarchiche, mettono in discussione la sostenibilità di queste teorie.
- **conseguenza naturale e probabile**: concerne i casi di in cui uno sviluppatore o un utente di non intendono né hanno conoscenza a priori di un reato.
  - La responsabilità è attribuita allo sviluppatore o all’utente se il danno è conseguenza naturale e probabile della loro condotta, esponendo gli altri in modo imprudente o negligente al rischio

3. Controllo del monitoraggio

Ci sono quattro meccanismi principali per affrontare il monitoraggio dei CIA.

1. Elaborare predittori dei utilizzando la conoscenza del dominio
2. Utilizzare la simulazione sociale per scoprire schemi ricorrenti di criminalità
3. Affrontare la tracciabilità lasciando indizi rivelatori nelle componenti che costituiscono gli strumenti dei CIA
4. Effettuare monitoraggio intersistemico e avvalersi dell’auto-organizzazione tra sistemi
   - concepire un sistema che assume il ruolo di paziente morale.

4. Affrontare la psicologia

> Ci sono due preoccupazioni principali relative all’elemento psicologico dei CIA: la manipolazione degli utenti e (nel caso dell’ antropomorfica) la creazione in un utente del desiderio di compiere un crimine.

Se gli antropomorci costituiscono un problema, allora possono esserci due approcci:

- limitare gli AA antropomorfici che consentono di simulare un crimine
- servirsi di antropomorfici come modo per respingere i reati sessuali simulati (incompatibile con il primo)

9. Uso dell'IA per il bene sociale

L'idea di AI per il bene sociale

L’idea di “intelligenza articiale per il bene sociale” (d’ora in poi) sta diventando popolare in molte società dell’informazione e sta guadagnando terreno nella comunità di AI.

Pressoché quotidianamente, infatti, compaiono nuove applicazioni di AI for Social Good (AI4SG), che rendono possibile e facilitano il raggiungimento di risultati socialmente positivi prima irrealizzabili, inaccessibili o semplicemente meno fattibili in termini di efficienza ed efficacia.

Chiaramente, le metriche esistenti, come la redditività o la produttività commerciale, misurano bene la domanda nel mondo reale, ma rimangono inadeguate.
L’AI deve essere valutata rispetto a **risultati socialmente validi**.

Arontare l’AI4SG *ad hoc*, analizzando aree di applicazione speciche, è indice della presenza di un fenomeno, ma non lo spiega, né suggerisce come altre soluzioni di potrebbero e dovrebbero essere disegnate per sfruttare appieno il potenziale dell’AI.
Inoltre, molti progetti che generano risultati socialmente buoni avvalendosi dell’AI non si (auto)descrivono in questi termini.

Tali carenze sollevano almeno due rischi principali: **fallimenti imprevisti** e **opportunità mancate**.

> **Fallimenti imprevisti**
>
> Come qualsiasi altra tecnologia, le soluzioni di sono modellate da valori umani. Tali valori, se non sono accuratamente selezionati e promossi, possono generare scenari di “AI buona andata storta”.
>
> L’AI può “fare più male che bene”, laddove amplica invece di mitigare i mali della società, per esempio ampliando anziché restringendo le disuguaglianze esistenti o esacerbando i problemi ambientali.

Opportunità perse

> Risultati socialmente buoni dell’ possono in realtà sorgere in modo del tutto accidentale, per esempio attraverso l’applicazione fortuita di una soluzione di in un contesto diverso.
>
> Per ogni “successo accidentale”, ci possono essere innumerevoli esempi di opportunità mancate per sfruttare i benefici dell’AI nel promuovere risultati socialmente buoni in contesti diversi

Al fine di evitare fallimenti inutili e opportunità mancate, l’AI trarrebbe vantaggio da un’analisi dei **fattori essenziali che supportano e assicurano il design e l’implementazione di AI di successo**:

1. falsicabilità e implementazione incrementale;
2. garanzie contro la manipolazione dei predittori;
3. intervento contestualizzato in ragione del destinatario;
4. spiegazione contestualizzata in ragione del destinatario e finalità trasparenti;
5. tutela della privacy e consenso dell'interessato;
6. equità concreta;
7. semantizzazione adatta all'umano

Una volta identicati questi fattori, le domande che possono formularsi sono a loro volta le seguenti:

- *in che modo questi fattori dovrebbero essere valutati e trattati?
- *da chi?*
- *con quale meccanismo di sostegno?*

Una definizione di AI4SG

Un progetto di ha successo nella misura in cui contribuisce a ridurre, mitigare o eliminare un determinato problema sociale o ambientale, senza introdurre nuovi danni o amplicare quelli esistenti.

> AI4SG = def. il design, lo sviluppo e l'implementazione di sistemi di in modo da (i) prevenire, mitigare o risolvere i problemi che incidono negativamente sulla vita umana e/o sul benessere del mondo naturale e/o (ii) consentire sviluppi preferibili dal punto di vista sociale e/o sostenibili dal punto di vista ambientale

1. Falsicabilità e implementazione incrementale

L'affidabilità è essenziale affinché la tecnologia in generale e le applicazioni di AI in particolare siano adottate e abbiano un signicativo impatto positivo sulla vita umana e sul benessere ambientale.

Sebbene non esistano regole o linee guida universali che possano assicurare o garantire l'affidabilità, la **falsicabilità** è un fattore cruciale per migliorare l'affidabilità delle applicazioni tecnologiche.

> La falsicabilità implica la specificazione, e la possibilità di verica empirica, di uno o più requisiti critici, cioè di una condizione, risorsa o mezzo necessari anché una capacità sia pienamente operativa, di modo tale che qualcosa non potrebbe o dovrebbe funzionare senza di essa

La sicurezza è un requisito critico ovvio. Dunque, affinché un sistema di sia affdabile, la sua sicurezza dovrebbe essere falsicabile.

I requisiti critici dovrebbero essere testati con un ciclo di implementazione incrementale.
Effetti pericolosi inintenzionali possono manifestarsi solo a seguito dei test.
I test possono essere eseguiti:

- con prove formali (difficili)
- nel mondo reale, se è sicuro farlo
- in **simulazioni**, che consentono di vericare se i requisiti critici (pensiamo di nuovo alla sicurezza) sono soddisfatti in base a una serie di ipotesi formali

Dall'analisi precedente discende che il fattore essenziale di falsicabilità e di implementazione incrementale comprende un ciclo:

1. requisiti ingegneristici falsicabili (cosicché sia almeno possibile sapere se i requisiti non sono soddisfatti);
2. test di falsificazione per migliorare progressivamente i livelli di affidabilità;
3. correzione delle ipotesi a priori;
4. allora e soltanto allora implementazione in un contesto sempre più ampio e critico.

> ***1. I progettisti di dovrebbero identicare i requisiti falsicabili e testarli in fasi incrementali dal laboratorio al "mondo esterno".***

2. Garanzie contro la manipolazione dei predittori

> Il potere predittivo dell'AI affronta due rischi: la manipolazione dei dati di input e l'eccessiva dipendenza da indicatori non causali.

Manipolazione dei dati di input

Quando il modello utilizzato è facile da comprendere "sul campo", si presta ad abusi o "manipolazioni", indipendentemente dal fatto che sia utilizzata l'AI.
L'introduzione dell' complica le cose, a causa della dimensione a cui l' viene di regola applicata.

> Se sono note le informazioni utilizzate per prevedere un dato risultato, un agente con tali informazioni (che si prevede intraprenderà una determinata azione) può modicare il valore di ciascuna variabile predittiva per evitare un intervento

Eccessiva dipendenza da indicatori non causali

Al contempo, c'è il rischio che un'eccessiva dipendenza da indicatori non causali – cioè dati che sono correlati con, ma non causa di, un fenomeno – possa distogliere l'attenzione dal contesto in cui il designer di AI4SG sta cercando di intervenire.

> Per essere efficace, qualsiasi intervento di questo tipo dovrebbe modicare le cause alla base di un dato problema piuttosto che i predittori non causali.

> ***2. I designer di AI4SG dovrebbero adottare garanzie che***
> ***(i) assicurino che gli indicatori non causali non distorcano in modo inappropriato gli interventi e***
> ***(ii) limitino, quando appropriato, la conoscenza di come gli input inuenzano gli output dei sistemi di , per prevenire la manipolazione.***

3. Intervento contestualizzato in ragione del destinatario

È essenziale che il software intervenga nella vita degli utenti solo in modi rispettosi della loro autonomia.

L'attenzione prestata al bilanciamento è comune per le iniziative di AI4SG.

Il rischio di **falsi positivi** (intervento non necessario, creazione di disillusione) è spesso altrettanto problematico dei **falsi negativi** (nessun intervento dove necessario, limitazione dell'efficacia).**

Per questo, un adeguato intervento *contestualizzato in ragione del destinatario* è quello che raggiunge il giusto livello di perturbazione, rispettando al contempo l'autonomia tramite le opzioni che offre.

> ***3. I designer di dovrebbero costruire sistemi decisionali in dialogo con gli utenti che interagiscono con questi sistemi e ne sono inuenzati; sulla base della comprensione delle caratteristiche degli utenti, delle modalità di coordinamento, delle nalità e degli eetti di un intervento; e nel rispetto del diritto degli utenti di ignorare o modicare gli interventi.***

4. Spiegazione contestualizzata in ragione del destinatario e finalità trasparenti

> Le applicazioni di dovrebbero essere disegnate in modo tale da rendere spiegabili le operazioni e i risultati di tali sistemi e trasparenti i loro scopi.

Rendere spiegabili i sistemi di è un importante principio etico.
La **spiegazione** di un intervento dovrebbe essere contestualizzata in modo tale da risultare adeguata e tutelare l'autonomia del destinatario.

Il **livello di astrazione** (LdA), dipende da cosa viene spiegato, a chi e per quale scopo.
Un LdA è un elemento chiave di una teoria e dunque di ogni spiegazione.
Una teoria comprende cinque elementi costitutivi:

1. un sistema
2. uno scopo
3. un livello di astrazione
4. un modello
5. una struttura del sistema

Il LdA fornisce la concettualizzazione del sistema. In ragione dello scopo e della sua granularità, non tutti i LdA sono appropriati per un dato destinatario.

Anche la trasparenza sull'obiettivo del sistema (cioè lo scopo del sistema) è cruciale, poiché deriva direttamente dal principio di autonomia.
Rendere trasparenti gli obiettivi e le motivazioni degli stessi sviluppatori di è un fattore cruciale per il successo di qualsiasi progetto, ma può contrastare con lo scopo stesso del sistema. Ecco perché è fondamentale valutare, in fase di design, qual è il livello di trasparenza (ossia quanta

trasparenza, di che tipo, per chi e su cosa) che il progetto adotterà, dato il suo obiettivo generale e il contesto di implementazione.

> ***4. I designer di dovrebbero scegliere un livello di astrazione per la spiegazione dell' che soddis lo scopo esplicativo auspicato e sia appropriato al sistema e ai destinatari; quindi dovrebbero fornire argomenti che siano razionalmente e adeguatamente persuasivi anché i destinatari forniscano la spiegazione; e assicurare che l'obiettivo (lo scopo del sistema) per cui viene sviluppato e implementato un sistema di sia conoscibile per impostazione predenita ai destinatari dei suoi risultati.***

5. Tutela della privacy e consenso dell'interessato

> La privacy è considerata **una condizione essenziale per la sicurezza e la coesione sociali**.

In circostanze in cui l'urgenza non è così pressante, è possibile ottenere il previo consenso di un soggetto all'utilizzo dei suoi dati. Il livello o il tipo di consenso richiesto può variare in ragione del contesto.

Tuttavia, è possibile trovare un equilibrio tra il rispetto della privacy del paziente e la creazione di un'AI4SG efficace:
- anonimizzare i dati

> ***5. I designer di devono rispettare la soglia di consenso stabilita per il trattamento delle raccolte di dati personali.***

6. Equità concreta

Gli sviluppatori di si adano di regola ai dati, che possono essere distorti in modo tale da avere eetti socialmente rilevanti. Tale pregiudizio (**bias**) può estendersi al processo decisionale algoritmico che è alla base di molti sistemi di AI, con conseguenze che sono inique per i soggetti del processo decisionale e, pertanto, possono violare il principio di giustizia.

Le iniziative di che si basano su dati distorti possono propagare tale pregiudizio attraverso un circolo vizioso. Questo ciclo inizierebbe con un insieme di dati distorto che informa una prima fase del processo decisionale dell', con conseguenti azioni discriminatorie, che a loro volta portano alla raccolta e all'uso di dati distorti.

Chiaramente, i designer devono sterilizzare gli insiemi di dati adoperati per addestrare l'AI.

> ***6. I designer di dovrebbero rimuovere dagli insiemi di dati rilevanti le variabili e i proxy che sono irrilevanti per un risultato, tranne nel caso in cui la loro introduzione supporti inclusione, sicurezza o altri imperativi etici.***

7. Semantizzazione adatta all'umano

L'AI deve consentire agli esseri umani di curare e promuovere il proprio "capitale semantico", ovvero

> qualsiasi contenuto che può incrementare il potere di qualcuno di dare signicato e conferire senso a (**semantizzare**) qualcosa.

Abbiamo spesso la capacità tecnica di automatizzare la creazione di signicato e senso (semantizzazione) tramite l', ma possono anche manifestarsi sducia o ingiustizia se lo facciamo con noncuranza. Da ciò emergono due problemi.

Il primo problema è che il soware di può denire la semantizzazione in modo divergente dalle nostre scelte.
Il secondo problema consiste nel fatto che, in un contesto sociale, sarebbe inattuabile per il soware di denire tutti i signicati e i sensi. La semantizzazione è in una certa misura soggettiva, perché chi o che cosa è coinvolto nella semantizzazione è anche in parte costitutivo del processo e del suo esito.

> La soluzione a questi due problemi si basa sulla distinzione tra i compiti che dovrebbero o non dovrebbero essere delegati a un sistema articiale. L' dovrebbe essere impiegata per facilitare la semantizzazione adatta all'umano, ma non per fornirla di per sé.

> ***7. I designer di AI non dovrebbero ostacolare la capacità delle persone di semantizzare (cioè di dare signicato e conferire senso a) qualcosa***

Fattori	**Migliori pratiche**	**Principi etici**
Falsificabilità e implementazione incrementale	Identificare i requisiti falsificabili e testarli in fasi incrementali dal laboratorio al “mondo esterno”.	Beneficenza Non maleficenza
Garanzie contro la manipolazione dei predittori	Adottare garanzie che (i) assicurino che gli indicatori non causali non distorcano in modo inappropriato gli interventi e (ii) limitino, quando appropriato, la conoscenza di come gli input influenzano gli output dei sistemi di AI4SG, per prevenire la manipolazione.	Beneficenza Non maleficenza
Intervento contestualizzato in ragione del destinatario	Costruire sistemi decisionali in dialogo con gli utenti che interagiscono con questi sistemi e ne sono influenzati; sulla base della comprensione delle caratteristiche degli utenti, delle modalità di coordinamento, delle finalità e degli effetti di un intervento; e nel rispetto del diritto degli utenti di ignorare o modificare gli interventi.	Beneficenza Autonomia
Spiegazione contestualizzata in ragione del destinatario e finalità trasparenti	Scegliere un livello di astrazione per la spiegazione dell’IA che soddisfi lo scopo esplicativo auspicato e sia appropriato al sistema e ai destinatari; quindi fornire argomenti che siano razionalmente e adeguatamente persuasivi affinché i destinatari forniscano la spiegazione; e assicurare che l’obiettivo (lo scopo del sistema) per cui viene sviluppato e implementato un sistema di AI4SG sia conoscibile per impostazione predefinita ai destinatari dei suoi risultati.	Beneficenza Spiegabilità
Tutela della privacy e consenso dell’interessato	Rispettare la soglia di consenso stabilita per il trattamento delle raccolte di dati personali.	Beneficenza Autonomia Non maleficenza
Equità concreta	Rimuovere dagli insiemi di dati rilevanti le variabili e i proxy irrilevanti per un risultato, tranne nel caso in cui la loro introduzione supporti inclusione, sicurezza o altri imperativi etici.	Beneficenza Giustizia
Semantizzazione adatta all’umano	Non ostacolare la capacità delle persone di semantizzare (cioè di dare significato e conferire senso a) qualcosa.	Beneficenza Autonomia

100. Glossario

epistemologicamente

Che concerne l’epistemologia, cioè la filosofia della scienza, e in senso più ampio la conoscenza dei metodi delle scienze e dei principî secondo i quali la scienza costruisce sé stessa: *dottrina*, teoria

Governance

L'insieme dei principi, delle regole e delle procedure che riguardano la gestione e il governo di una società, di un'istituzione, di un fenomeno collettivo.

Infosfera

Con il termine infosfera nella filosofia dell'informazione si intende la globalità dello spazio delle informazioni. Pertanto l'infosfera include sia il ciberspazio (Internet, telecomunicazioni digitali) sia i mass media classici.

ontologicamente

Che riguarda la conoscenza dell’essere, della realtà, dell’oggetto in sé