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 [[4. Un quadro unificato di principi etici per l'IA#^828f30|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 [[4. Un quadro unificato di principi etici per l'IA#^57402f|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 [[4. Un quadro unificato di principi etici per l'IA#^828f30|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 (**[[7. La mappatura dell'etica degli algoritmi#^6b736e|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 [[4. Un quadro unificato di principi etici per l'IA#^6f8295|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

![[Pasted image 20240119191541.png]]