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 (*I Principi di Asilomar per l'IA, La Dichiarazione di Montréal per l'IA responsabile, Dichiarazione su intelligenza artificiale, robotica e sistemi autonomi*) 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**.

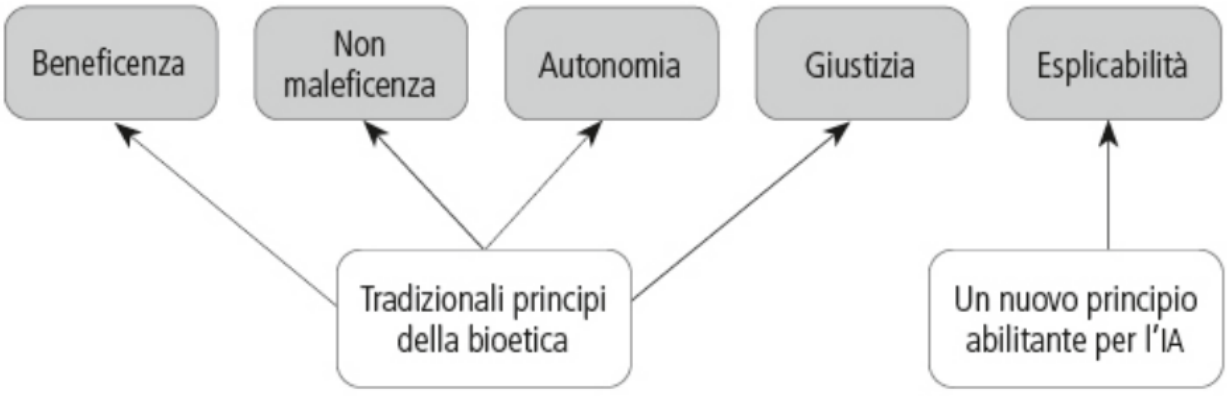

Figura 4.1 Un quadro etico dei cinque principi fondamentali per l'IA.

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.

6. Precauzione

> Il principio di precauzione delinea la necessità di un atteggiamento di cautela intesa come anticipazione preventiva del rischio di fronte all’incertezza epistemologica del sapere scientifico

presuppone una presa d’atto della relatività del sapere scientifico e della incapacità costitutiva della scienza e della tecnologia di offrire risultati certi