Etica e intelligenza artificiale: i principi di Asilomar e la Human-centered AI

In una recente intervista con il filosofo della scienza Michele Marsonet ho affrontato il tema dell'intelligenza artificiale: essa rappresenta «il coronamento di un vecchio sogno dell'umanità, quello di riprodurre e meccanizzare il processo del pensiero». Naturalmente, accanto alle enormi possibilità che apre l'AI, vi sono anche tanti rischi: «occorre affrontare la questione – concludeva Marsonet – elaborando dei codici etici in grado di regolamentare l'utilizzazione e lo sviluppo dei sistemi artificiali. (vedi Filosofia, scienza, intelligenza artificiale. A colloquio con Michele Marsonet).

Un argomento, dunque, di interesse pluridisciplinare.

La conferenza di Dartmouth (1956) è passata alla storia come l'incontro che ha segnato la nascita del campo di ricerca sull'intelligenza artificiale. Si è trattato di un incontro fatto per aprire un nuovo spazio di ricerca «sulla base della congettura per cui, in linea di principio, ogni aspetto dell'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano essere descritte così precisamente da poter costruire una macchina che le simuli» (Proposta di Dartmouth, p. 1). Da questo evento, numerosi passi in avanti sono stati fatti e, assieme ai numerosi successi e alle nuove potenzialità, è andata sempre più crescendo la necessità di elaborare degli orientamenti, di elaborare una "guida etica".

Uno dei frutti di questa necessità sono i 23 principi di Asilomar (stilati nel gennaio del 2017). Si tratta di un testo suddiviso in tre aree: la prima sulla "ricerca", la seconda su "etica e valori", la terza e ultima sui "problemi di scenario". Questi principi sono stati elaborati proprio per guidare la ricerca verso uno sviluppo dell'AI «benefico e sicuro»: essi riguardano, difatti, temi come la «trasparenza della ricerca», la «responsabilità», il «controllo umano sui sistemi di AI». I «i sistemi di AI autonomi – si legge al numero 10 – dovranno essere progettati in modo che i loro obiettivi e comportamenti possano essere allineati con i nostri valori per tutto l'arco del loro esercizio» e «la super-intelligenza deve essere sviluppata solo al servizio di ideali etici ampiamente condivisi e per il beneficio di tutta l'umanità, non di uno Stato o di un'organizzazione» (n. 23).

Questo vuol dire che il "problema AI" va affrontato – come dicevo inizialmente – in modo pluridisciplinare, unico capace di garantire una base solida per uno sviluppo sicuro o – come dice Fei-Fei Li – per un approccio di «IA centrata sull'uomo»: «At Stanford

HAI – dice Fei-Fei Li (direttrice dell'Istituto) – our vision is led by our commitment to studying, guiding and developing human-centered AI technologies and applications».

Una vera sfida che chiama in causa altre importanti questioni come l'oggettività dei valori e la natura dell'uomo, presupposti fondamentali per raggiungere l'«ampia condivisione» di cui si parlava.

Giovanni Covino
bricioledifilosofia.com