Ordine internazionale e diritti umani
International Legal Order and Human Rights
Ordenamiento Jurìdico Internacional y Derechos Humanos
Ordre Juridique International et Droits de l'Homme
Diretta da Claudio Zanghì, Lina Panella, Carlo Curti Gialdino
EDITORIALE SCIENTIFICA

OSSERVATORIO SUL CONSIGLIO DEI DIRITTI UMANI E SUI COMITATI ONU N. 5/2024

1. ALGORETICA E SOFT STANDARDS: IL CONTRIBUTO DELLA UN HUMAN RIGHTS MACHINERY AL DIBATTITO INTERNAZIONALE SULLE SFIDE POSTE DALLE NUOVE TECNOLOGIE E DALL'INTELLIGENZA ARTIFICIALE

1. *Osservazioni introduttive per una definizione dell'algoretica*

La sfida complessa posta dal rapido evolversi delle conoscenze e degli strumenti operativi di natura tecnologica ai fini di una appropriata regolamentazione che tuteli la persona e i diritti e le libertà di cui è titolare nel contesto comunitario di appartenenza è sempre più centrale nel dibattito internazionale.

Essa muove dalla necessità di declinare alcune nozioni prettamente tecniche nel linguaggio giuridico rappresentato dal diritto internazionale dei diritti umani e, in tale prospettiva, la Human Rights Machinery di Ginevra ha proposto percorsi di approfondimento e occasioni di incontro e dialogo sul tema coinvolgendo tutti gli attori, pubblici e privati, interessati a fornire il proprio contributo.

In questo contributo, pertanto, è indispensabile muovere dai concetti a fondamento della discussione in atto, al fine di delineare alcune considerazioni preliminari formulate dai predetti attori, che informeranno il progressivo posizionamento della Machinery sul tema, in un ambito di comune interesse ed esercizio, ovvero l'intelligenza artificiale generativa.

Quando si analizzano le modalità operative di matrice tecnologica, il primo concetto prioritario è rappresentato dall'algoritmo: la funzione consta di tre componenti, ovvero l'input, il processo di elaborazione, e l'output. La prima si sostanzia nell'insieme di dati che l'algoritmo recepisce e sui quali opera, dal singolo numero al set complesso di dati quantitativi. Il secondo si articola nella sequenza di operazioni proprie dell'algoritmo, impostate e tuttavia automatizzate anche in ragione del volume massivo di dati lavorati. La terza costituisce il risultato finale del processo operativo basato sull'algoritmo. La chiarezza, la rapidità e la precisione dell'algoritmo contribuiscono al raggiungimento di un output particolarmente soddisfacente. Allorché l'algoritmo è utilizzato nei sistemi che ricorrono all'intelligenza artificiale, esso viene ad assumere una connotazione strutturale ed operativa tipica: non soltanto opera sul dato in maniera tradizionale, bensì è il dato stesso – singolo numero o set complesso di dati – a correggere progressivamente l'algoritmo per il suo migliore funzionamento e la sua adattabilità al contesto operativo: in quest'ultimo, infatti, l'algoritmo è strumentale per la raccolta e classificazione dei dati, per la individuazione ed il riconoscimento di modelli ricorrenti basati sui dati, per fornire orientamenti decisionali, per formulare considerazioni predittive.

In questa accezione complessa, il tentativo di rendere l'algoritmo una funzione completamente oggettiva, seppur originata dal ragionamento della mente umana, è uno tra i rischi emergenti nel quadro dello sviluppo marcato delle nuove tecnologie, affrontato attraverso il contributo della dottrina per la creazione di una disciplina circostanziata nel quadro dell'etica: l'algoretica.

Essa studia la concettualizzazione tecnica e l'utilizzo dell'algoritmo a partire dai parametri quantitativi e qualitativi, dalle variabili conosciute o presunte, dagli eventuali errori operativi che dipendono da chi concepisce e opera sull'algoritmo: in questa declinazione, l'algoritmo è oggetto d'indagine nella prospettiva soggettiva, in grado di determinare una lettura del dato non equilibrata, potenzialmente discriminatoria, capace di manipolare le attitudini, le preferenze, le opinioni del singolo in modo esponenzialmente ampio nello spazio digitale aperto. Dunque, l'algoretica si focalizza su una lettura positiva del funzionamento dell'algoritmo, in linea con il rispetto dei principi e la protezione dei diritti umani, come ben esplicitato dal Prof. Benanti: "*algorethics seek to encode ethical principles into the software so that algorithmic decisions can avoid harmful or undesired consequences*" (B. D. MITTELSTADT, et al., *The Ethics of Algorithms: Mapping the Debate,* in *Big Data & Society*, 3(2), 2016, pp. 1-21; P. BENANTI, *The Urgency of an Algorethics*, in Discov *Artificial Intelligence*, 3 (11), 2023; P. SCHERZ, *AI as Person, Paradigm, and Structure: Notes toward an Ethics of AI*, in *Theological Studies*, 85(1), 2024, pp. 124-144).

In tale ottica, la disciplina intende approfondire e dare riscontro ad alcuni quesiti-chiave che rilevano sotto il duplice profilo tecnico ed etico, anche in previsione della formulazione di possibili standard giuridici, *soft* e *hard,* atti a regolamentare il corretto utilizzo dell'algoritmo nel prossimo futuro tecnologico. L'intelligenza artificiale può operare in modo tale da rispettare la dignità e tutelare i diritti umani, ed è percorribile una configurazione tecnica atta a prevenirne l'utilizzo con impatto chiaramente discriminatorio nei confronti di individui o gruppi di individui? Qualora l'impatto operativo sia lesivo dei diritti, la vittima ha possibilità di rivendicarne la responsabilità a carico di chi ha configurato e reso accessibile un servizio basato sull'intelligenza artificiale? E, ancora, secondo quali modalità i soggetti produttori e distributori di tale tipo di servizi riescono ad assicurarne un funzionamento sicuro ed affidabile nonché chiaro e trasparente?

Dalla formulazione di tali quesiti sono derivate alcune linee di riflessione, promosse primariamente su scala internazionale (si veda la 2020 "*Call for AI Ethics*" firmata dalla Pontificia Accademia per la Vita e dal Ministero dell'Innovazione insieme ad un ampio numero di enti privati del settore tech al fine di "*promote an ethical approach to artificial intelligence, giving shared responsibility among international organisations, governments, institutions and companies, to mould a future of digital innovation at the service of mankind, individually and as a whole*", aggiornata in sede UNESCO nel formato della Raccomandazione adottata per acclamazione dai 193 Stati membri il 23 novembre 2021), per la individuazione di alcuni principi fondamentali propri dell'algoretica, al fine di affrontare in modo propositivo e dinamico le sfide insite nel rapido sviluppo ed applicazione delle conoscenze tecnologiche, ivi inclusa l'intelligenza artificiale.

Innanzitutto, i sistemi basati sull'intelligenza artificiale devono essere 'explainable', ovvero indicare con estrema chiarezza quale informazione è prodotta attraverso l'utilizzo della macchina, lasciando la persona libera di decidere se farne uso. In caso positivo, l'uso stesso deve essere funzionale al miglioramento delle relazioni interpersonali e sociali e al potenziamento delle capacità personali per la realizzazione di determinate attività. I creatori e produttori di detti sistemi devono assumersi una precisa responsabilità nell'avviare e condurre il processo di sviluppo tecnologico mirato al funzionamento di un sistema basato

sull'intelligenza artificiale, in tale accezione non potendosi attribuire altra responsabilità alla sola macchina attraverso un percorso di umanizzazione di quest'ultima. Nel funzionamento dei sistemi in parola, che non presenta una connotazione umana o morale, è essenziale assicurare la piena imparzialità dei processi di raccolta ed analisi del dato, preservando pertanto la macchina dall'assumere un atteggiamento soggettivo e potenzialmente discriminatorio. Il principio della 'reliability' garantisce altresì che il sistema basato sull'intelligenza artificiale non crei situazioni di rischio e/o in danno della persona, ed infine ne metta in sicurezza il dato proteggendolo in termini di riservatezza.

La risposta ai quesiti sopra richiamati, unitamente alla formulazione dei principi fondamentali ha permesso di delineare il portato attuale dell'algoretica in numerose aree chiave. Nel settore sanitario l'intelligenza artificiale contribuisce per la raccolta ed analisi del dato del paziente e, al contempo, ne rispetta la riservatezza sulla base del consenso informato al trattamento del dato medesimo in favore della conoscenza e dello scambio delle informazioni per il benessere e la salute collettiva. Nel contesto educativo, l'algoretica incentiva i percorsi di apprendimento delle giovani generazioni, ancorché in modo personalizzato ove necessario, motivando la creatività ed il pensiero critico. Nel settore finanziario, l'algoretica può promuovere condizioni di stabilità e più ampio accesso ai servizi, prevenendo fenomeno di matrice corruttiva o criminosa.

È evidente che, in costanza delle dinamiche evolutive insite nelle tecnologie, la disciplina algoretica sarà ulteriormente definita allo scopo di far fronte a criticità dipese da fattori imprevisti e non conosciuti, dalla disuniformità delle risposte che non consentono di compilare standard etici comuni per il corretto funzionamento dei sistemi basati sull'intelligenza artificiale, dalla tensione volta al superamento di nuovi dilemmi concettuali (riservatezza-sicurezza; accuratezza-equità; efficienza-chiarezza operativa) nella lettura delle sfide tecnologiche contemporanee.

2. L'intelligenza artificiale generativa e i rischi per la protezione dei diritti umani

La formulazione dei principi a fondamento dell'algoretica ha agevolato un interessante dibattito nel quadro sistemico onusiano focalizzato sulle dinamiche dello sviluppo tecnologico dell'intelligenza artificiale con impatto specifico sul livello di protezione dei diritti umani (M. LATONERO, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, in *Data & Society*, 2018; J. FJELD et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society. 2020).

In tale prospettiva, il più elevato rischio di violazione delle fattispecie giuridiche di titolarità individuale e collettiva deriva, oggi, dalla c.d. intelligenza artificiale generativa: il funzionamento della macchina, pur su impulso umano, è in grado di produrre contenuti autonomi nella forma di testo, immagine, video, suono, codice, rappresentazione grafica tridimensionale, a partire dai dati già disponibili sulla rete. In altre parole, l'algoritmo riesce a catturare ogni dato utile e si auto-alimenta, generando risultati che simulano il ragionamento umano anche nella prassi c.d. predittiva.

In occasione del suo intervento nei lavori del Summit dedicato a "*Generative Artificial Intelligence and Human Rights*", l'Alto Commissario per i Diritti Umani ha sottolineato con forza il paradosso insito in tale metodologia operativa dei sistemi basati sull'intelligenza artificiale: ciò richiede una riflessione attualizzata che muova dal rinnovamento della tassonomia inerente i rischi prodotti dall'intelligenza artificiale generativa compressivi della tutela dei

diritti umani, e la tassonomia deve indiscutibilmente informare la governance internazionale e nazionale, includendo in tale processo anche e soprattutto gli attori privati che investono nel settore tecnologico: "*Generative AI is not a local or national phenomenon. It will have impact on everyone – and it demands a global, collaborative approach. We need to make sure that protecting people's rights is at the centre of that approach. This requires not just dialogue, but action – action that draws upon the collective wisdom and guidance of established frameworks. Our collaboration should unite States, corporations, civil society, and individuals in a shared mission: to ensure that AI serves humanity's best interests, co-creating a world in which technology does not just serve the interests of the wealthy and powerful, but enables universal advancement of human dignity and rights*".

Le considerazioni dell'Alto Commissario confermano il peso specifico del diritto internazionale dei diritti umani quale standard che, nella sua duplice portata positiva e vincolante nonché morale e non scritta, rappresenta il quadro di riferimento per una governance globale dell'intelligenza artificiale e delle tecnologie, e non può e non deve essere compromesso dalla possibile risposta in termini di regolamentazione frammentata e disarmonica percorsa al livello nazionale e regionale.

In particolare, l'intelligenza artificiale generativa è un ambito operativo sfidante per la protezione dei diritti umani, associata all'utilizzo delle stesse tecnologie che hanno posto e pongono rischi compressivi dei diritti e delle libertà stesse. A titolo esemplificativo, l'utilizzo alterato di dati per la creazione di una informazione non veritiera, che dunque alimenta la "*online misinformation and disinformation*", è oramai una condizione diffusa sulla rete; parimenti, la raccolta e l'uso di dati personali di evidente sensibilità da parte di operatori privati, alimentati peraltro dallo stesso titolare del dato proprio ricorrendo all'intelligenza artificiale creativa, limita in modo preoccupante la riservatezza dell'individuo ed utente digitale. Ancora, il funzionamento dei sistemi basati sull'intelligenza artificiale creativa incoraggia una interazione tra la persona e la macchina che, consapevolmente o inconsapevolmente, orienta il pensiero e la sua manifestazione limitando la libertà di opinione in modo fattuale.

Tali esempi, considerati nella loro dimensione singola, facilitati da un'intelligenza artificiale creativa che opera in diversi ambiti e che dunque presenta una connotazione multi-sistema, potrebbero condurre a molteplici e contestuali rischi per la protezione dei diritti umani: dunque, la discussione in atto nella *Human Rights Machinery* ginevrina non si articola soltanto sulla opportunità di identificare soluzioni comuni nel breve e medio termine, bensì soprattutto nel lungo periodo e comunque coinvolgendo in modo inclusivo – come occorso nel quadro della Presidenza italiana del G7 per il 2024 con riferimento, ad esempio, all'iniziativa G7 Toolkit for AI in the Public Sector - tutti gli attori pubblici e privati, co-responsabili nei processi di sviluppo tecnologico governati in funzione del consolidamento degli standard giuridici propri del diritto internazionale dei diritti umani.

3. Il ruolo delle tech companies e l'attuazione dei Principi Guida delle Nazioni Unite su Impresa e Diritti Umani nella prospettiva dello sviluppo tecnologico accelerato: quali le misure di regolamentazione ottimali?

Nel suo intervento al summenzionato Summit, l'Alto Commissario per i Diritti Umani ha posto l'accento sul ruolo determinante degli attori privati che operano nel settore d'impresa correlato alle tecnologie e allo spazio digitale.

Invero, è in questo ambito che l'Ufficio dell'Alto Commissario ha concepito e sta realizzando una progettualità mirata. Il B-Tech Project, avviato nel 2019 sotto la guida di Shift e la partecipazione della *Global Network Initiative* (M. A. SAMWAY, *The Global Network Initiative: How Can Companies in the Information and Communications Technology Industry Respect*

Human Rights, in D. BAUMANN-PAULY, J. NOLAN (eds.), *Business and Human Rights: From Principles to Practice*, New York: Routledge, 2016, pp. 136-146), intende promuovere il dibattito internazionale lungo le direttrici della ricerca, dello scambio di informazioni, della realizzazione di pratiche operative che incentivino lo sviluppo tecnologico senza incrementare il rischio di violazione dei diritti umani attraverso i sistemi basati sull'intelligenza artificiale creativa.

Il progetto concentra l'attenzione sul ruolo dell'impresa nella configurazione, realizzazione, produzione e distribuzione di prodotti tecnologici che utilizzano l'intelligenza artificiale creativa affrontando in modo preventivo e fornendo soluzioni operative a fronte del rischio o della comprovata violazione dei diritti umani (M. FLYVERBOM, R. DEIBERT, D. MATTEN, *The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business*, in *Business & Society*, 58(1), 2019, pp. 3-19.; I EBERT, A. BEDUSCHI, *Regulating Business Conduct in the Technology Sector: Gaps and Ways Forward in Applying the UNGPs*, Geneva Academy. April 2022; I. EBERT, *Fostering Business Respect for Human Rights in AI Governance and Beyond: A Compass for Policymakers to Align Tech Regulation with the UNGPs*, Carr Center for Human Rights Policy Harvard Kennedy School, Harvard University, 2024).

Affinché le *tech companies* improntino le rispettive strategie di sviluppo in tal senso, il presupposto primario risiede nell'adozione di un approccio produttivo basato sui diritti umani: ciò è possibile incorporando in modo comprensivo i Principi Guida delle Nazioni Unite su Impresa e Diritti Umani nelle politiche e nei processi d'impresa, in particolare quanto le *tech companies* realizzano prodotti e servizi basati sull'intelligenza artificiale creativa. Nel rispondere a tale esigenza, nel progetto sono state raccolte numerose buone pratiche che testimoniano l'impegno delle *tech companies* nell'adottare principi *ad hoc*, in linea con i predetti Principi Guida, in funzione di un utilizzo responsabile dell'intelligenza artificiale: la promozione dei valori e il controllo umano sulla macchina e sulla tecnologia in generale, la non discriminazione, la trasparenza, la chiarezza operativa, la responsabilità, la sicurezza, la protezione della riservatezza, la tutela dei diritti umani.

Al contempo le *tech companies* si sono adoperate in termini gestionali, con l'obiettivo di prevenire il rischio associato all'utilizzo dell'intelligenza artificiale creativa, formando i propri operatori specializzati e creando *team* multi-disciplinari nei quali la conoscenza più squisitamente tecnologica si coniuga con la necessità di introdurre nozioni articolate in ordine agli standard internazionali in materia di protezione dei diritti umani per le migliori pratiche di verifica e mitigazione del rischio derivante dall'intelligenza artificiale creativa. La prassi conferma che, al di là di una strutturazione formalizzata della governance d'impresa, i team diversamente denominati ("*responsible AI*," "*responsible innovation*", "*AI ethics*", "*AI safety*") operano in modo concreto e collaborativo o, altrimenti, si procede individuando una figura esterna particolarmente esperta e competente che affianca i team tecnici fornendo assistenza mirata proprio sull'adozione di un approccio basato sui diritti umani.

In ogni caso, ben oltre la soluzione strutturale ed operativa prescelta, rileva l'impatto concreto di tale approccio, a conferma dell'avvenuta ricezione e considerazione della dimensione garantista dei diritti umani in un processo produttivo di matrice tecnologica. Esso può esplicarsi in due principali modalità: la *disclosure* assicurata dall'impresa quanto all'adozione di una impostazione produttiva responsabile ricorrendo all'intelligenza artificiale creativa presso il pubblico in generale, e la creazione di un meccanismo rimediale che consenta alla persona eventualmente danneggiata per l'utilizzo dei prodotti o servizi dell'impresa di poter essere adeguatamente tutelata e risarcita.

La prima modalità è stata verificata nel progetto in parola raccogliendo numerose buone pratiche che sono similari nella formulazione dei principi basilari e nella gestione dei processi produttivi, mancando tuttavia una impostazione preventiva rispetto al possibile rischio associato all'utilizzo dell'intelligenza artificiale creativa e alla conseguente violazione dei diritti umani. Allo stesso tempo, sovente la *disclosure* ha presentato un linguaggio estremamente complesso in ragione della lettura tecnologica del prodotto, dunque non raggiungendo l'obiettivo della più ampia e chiara informazione in ordine al processo produttivo a vantaggio del pubblico, utenti e consumatori.

Sotto il profilo rimediale, emergono almeno due criticità per l'impresa che intende inserire nella sua filiera produttiva prodotti e servizi basati sull'intelligenza artificiale creativa. La prima attiene alla puntuale identificazione del nesso causale tra azione tecnologica e violazione dei diritti umani, mentre la seconda concerne la tipologia di responsabilità in capo all'impresa e dunque la determinazione – sostanziale e procedurale – della stessa misura rimediale. La prassi raccolta testimonia nuovamente la preferenza per un modello rimediale semplice e diretto, ovvero un canale di comunicazione creato dalla tech company per ricevere e gestire le segnalazioni degli utenti e risolvere il singolo caso, assicurandone la non reiterazione. Tuttavia, almeno in questa prospettiva, non si è ancora tenuta in considerazione l'ipotesi di violazione dei diritti umani in un ambiente multi-attoriale in cui la tipologia di co-responsabilità coinvolge attori tanto pubblici quanto privati e, pertanto, incide sulla possibile configurazione di un 'ecosistema rimediale' complesso.

Il progetto, poiché focalizzato sull'intelligenza artificiale creativa, non potrà non evolversi ulteriormente richiamando i principi summenzionati a fondamento dell'algoretica. Invero, la lettura della casistica in violazione dei diritti umani attraverso l'utilizzo di prodotti e servizi basati sull'intelligenza artificiale creativa risponde, oggi, ad una lettura etica per prevenire e riscontrare le esigenze di verifica e mitigazione del rischio. L'ampliamento dei principi dell'algoretica richiamando non soltanto gli standard giuridici internazionali ma anche, nel caso di specie, i Principi Guida delle Nazioni Unite su Impresa e Diritti Umani, è auspicabile, muovendo dal presupposto che etica, fiducia, sicurezza, diritti umani non sono elementi confliggenti bensì complementari in una strategia d'impresa ad impatto tecnologico.

Nella società attuale, reale e virtuale, il dato di fatto è l'estrema velocità delle conoscenze tecnologiche come il rapido evolversi della produzione di beni e servizi digitali da parte dell'impresa: ne discende una visione complessiva ed una co-responsabilità costruttiva da parte di tutti gli attori interessati da questo processo in divenire. Si tratta di delineare il percorso finalizzato alla creazione di una governance strutturale ed operativa per la configurazione, produzione e gestione dei predetti beni e servizi, con particolare riferimento a quelli che si basano sull'intelligenza artificiale creativa.

Se è vero che la componente attoriale d'impresa gioca un ruolo centrale in questo scenario, il peso specifico dei Principi Guida delle Nazioni Unite su Impresa e Diritti Umani è dunque confermato, già nelle parole di John Ruggie che li ha concepiti: "*(The UNGPs) are not merely a text. They were intended to help generate a new regulatory dynamic, one in which public and private governance systems, corporate as well as civil, each come to add distinct value, compensate for one another's weaknesses, and play mutually reinforcing roles—out of which a more comprehensive and effective global regime might evolve*".

Su base nazionale, le autorità governative sono dunque chiamate ad adottare uno 'smart-mix' di misure di regolamentazione, di orientamento, di garanzia in termini di trasparenza circa l'operato delle proprie imprese, rafforzando in tal modo la dimensione operativa della responsabilità ad esse attribuibile per ogni caso di violazione dei diritti umani,

lungo la catena di produzione e fornitura di beni e servizi tecnologici basati sull'intelligenza artificiale creativa. Al livello internazionale, documenti-guida quali gli stessi Principi o le Linee-guida dell'OCSE per le imprese multi-nazionali, possono rappresentare uno strumento utile per la definizione della stessa governance in materia di intelligenza artificiale generativa responsabile e per la identificazione di validi meccanismi rimediali di natura giudiziale e non, che contestualizzino in modo chiaro e definito la co-responsabilità pubblico-privato a fronte di violazioni dei diritti umani prodotte dall'intelligenza artificiale creativa.

CRISTIANA CARLETTI