Focus

AI: profili etici
Una prospettiva etica sull'Intelligenza Artificiale: princìpi, diritti e raccomandazioni

Stefano Quintarelli, Francesco Corea, Fabio Fossa, Andrea Loreggia, Salvatore Sapienza*

AN ETHICAL PERSPECTIVE ON ARTIFICIAL INTELLIGENCE: PRINCIPLES, RIGHTS AND RECOMMENDATIONS

ABSTRACT: As technologies become more and more pervasive in our everyday life new questions arise, for example, about security, accountability, fairness and ethics. These concerns are about all the realities that are involved or committed in designing, implementing, deploying and using the technology. This document addresses such concerns by presenting a set of practical obligations and recommendations for the development of applications and systems based on Artificial Intelligence (AI) techniques. These are derived from a definition of rights resulting from principles and ethical values rooted in the foundational charters of our social organization.

KEYWORDS: AI governance; security; accountably; fairness; ethical design

Nota metodologica

Il presente documento si prefigge lo scopo di definire un insieme di obblighi e raccomandazioni pratiche per lo sviluppo di applicazioni e sistemi basati su tecniche di Intelligenza Artificiale (IA).

* *Stefano Quintarelli, Associazione Copernicani. Mail: stefano@quintarelli.it; Francesco Corea, Università di Ca' Foscari. Mail: francesco.corea@unive.it; Fabio Fossa, Università di Torino. Mail: fabio.fossa@unito.it; Andrea Loreggia, Università di Padova. Mail: andrea.loreggia@unipd.it; Salvatore Sapienza, CIRSFID, Università di Bologna. Mail: salvatore.sapienza2@unibo.it.* *Il presente documento è stato elaborato dagli autori in piena autonomia, libertà ed indipendenza e riflette unicamente ed esclusivamente le opinioni del gruppo di lavoro. Il documento è stato redatto con il metodo del consenso ("consent"), ovvero con l'assenza di obiezioni significative per ogni suo enunciato. Pertanto, sebbene i membri del gruppo sostengano il documento nel suo complesso, non necessariamente essi condividono ogni singola affermazione contenuta nel documento stesso.*

Gli stessi sono derivati a partire da una definizione di diritti conseguenti a principi e valori etici radicati nei documenti fondamentali della nostra organizzazione sociale.
È stato redatto da un gruppo multidisciplinare di ricercatori costituito da Francesco Corea, Fabio Fossa, Andrea Loreggia, Salvatore Sapienza con la supervisione di Stefano Quintarelli.
Il manoscritto è stato quindi sottoposto alla revisione delle Proff. Maria Chiara Carrozza, Monica Palmirani, Francesca Rossi e del Prof. Carlo Casonato ed il testo è stato riesaminato alla luce dei commenti ricevuti.

1. Riferimenti

1.1. Lavori precedenti

Nella elaborazione si è tenuta in particolare considerazione quanto definito dai seguenti lavori:

- Partnership on AI (2016): lista di principi centrata sulla necessità di sviluppare una cultura di cooperazione fra ricercatori in IA, di garantire una distribuzione il più possibile equa dei benefici delle nuove tecnologie e il coinvolgimento di stakeholders pubblici e aziendali. Promosso dai principali *Over The Top*.
- Principi di Asilomar (2017): manifesto di principi di roboetica e linee guida per lo sviluppo delle nuove tecnologie, definiti da accademici e professionisti del settore.
- AI in the UK (2018): studio realizzato dalla Camera dei Lord a supporto della condivisione sociale dei benefici derivanti dall'uso di una IA trasparente e sicura.
- Villani (2018): report dell'esecutivo francese che definisce la propria strategia sull'Intelligenza Artificiale, elencando i principi fondamentali per il suo sviluppo.
- AI4People (2018): il documento, elaborato da Atomium-EISMD, chiarisce rischi e opportunità che l'IA presenta nei confronti della società contemporanea e delinea princìpi etici a cui adeguare ricerca e utilizzo dell'IA.
- CEPEJ (2018): il documento, redatto dalla European Commission for the Efficiency of Justice, ha lo scopo di valutare impatti etici e potenzialità dell'uso dell'IA in contesti giudiziari.
- HLEG_AI Ethics Guidelines (2019): linee guida definite dal gruppo di esperti della Commissione Europea per la creazione di una intelligenza artificiale affidabile ed attendibile.
- IEEE Ethically aligned design (2019): stabilisce che le tecnologie devono incorporare, attraverso pratiche da attuare già in sede di progettazione, i valori fondamentali a cui associare provvedimenti di policy e inquadramenti legali corrispondenti.

1.2. Fondamenti

Si è scelto di ancorare il lavoro ad alcune carte fondamentali del nostro tessuto sociale, partendo da quelle di carattere globale per arrivare al livello nazionale. Si fa quindi riferimento a:

- Sustainable Development Goals: formulati dalle Nazioni Unite nel 2015, con l'obiettivo di promuovere lo sviluppo sostenibile attraverso la soluzione di alcuni fra i maggiori problemi economico sociali dell'umanità; lo sviluppo dell'IA si lega a doppio filo con il loro raggiungimento.
  I guadagni di produttività offerti dell'IA devono anche favorire un'industrializzazione inclusiva rispettosa del lavoro umano, come stabilito negli obiettivi n.8 e n.9. Allo stesso tempo, bisogna

assicurare un'ampia diffusione dei suoi benefici al fine di ridurre le disuguaglianze, in accordo con l'obiettivo n. 10.

La promozione della parità di genere, fissata nell'obiettivo 5, richiede di garantire l'eliminazione di bias dal design degli algoritmi; la tutela di un'educazione paritaria e di qualità definita nell'obiettivo n. 4, si necessita della diffusione di cultura digitale a tutti i livelli d'istruzione.

Sul piano politico, la promozione di pace, giustizia ed istituzioni forti prevista nell'obiettivo 16 dipende in misura crescente da un utilizzo eticamente corretto dell'IA nella personalizzazione della comunicazione di massa, nella prevenzione e repressione del crimine e nell'amministrazione della giustizia, senza scadere in forme di manipolazione e di controllo statale autoritario. Per quanto riguarda la promozione della pace, in particolare, l'IA non deve in alcun modo sostituire il giudizio umano nei sistemi di arma.

- Dichiarazione Universale dei Diritti Umani: costituisce uno dei testi essenziali per ogni riflessione etico-giuridica da cui iniziare anche un dialogo sul quadro etico dell'IA. Tale dialogo è fondato sul riconoscimento della dignità umana quale fondamento del rispetto e della promozione dei diritti, indipendentemente dal sesso, dall'appartenenza etnica e religiosa, dalle opinioni politiche o da altri fattori che possono dar luogo a discriminazione. Il carattere universale dei diritti umani riconosciuti nella Dichiarazione la rende idonea a favorire una discussione globale e inclusiva sui temi e sulle sfide lanciate dall'IA nei confronti della società pluralista e multiculturale contemporanea. Di particolare rilevanza, per gli scopi di questo documento, sono la dignità (art. 1), la tutela della riservatezza (art. 12), la libertà di informazione (art. 19), e l'attenzione verso i temi di uguaglianza e non discriminazione, fondamentale nella riflessione sui *bias* algoritmici. Sono soggetti agli obblighi che scaturiscono dalla Dichiarazione gli Stati membri delle Nazioni Unite, che devono trasporne i suoi principi negli ordinamenti nazionali riguardanti i soggetti da loro regolati, non direttamente vincolati dalla Dichiarazione.
- Carta dei diritti fondamentali dell'Unione Europea: stabilisce valori e obiettivi fondamentali dell'Unione Europea e rappresenta un rilevante quadro di riferimento per lo sviluppo e l'uso di sistemi di IA. I valori delineati si incardinano sull'inclusione, la tolleranza, la giustizia, la solidarietà e la non discriminazione e si declinano nel rispetto della dignità e dei diritti umani, delle libertà individuali, degli ideali democratici, dell'uguaglianza dei cittadini davanti alla legge e dello stato di diritto. Alla luce di tali valori, L'UE si impegna a promuovere la pace e il benessere, la libertà e la sicurezza dei propri cittadini; a garantire giustizia e libertà di spostamento; a favorire lo sviluppo sostenibile basato su una crescita economica equilibrata e sulla stabilità dei prezzi, su un'economia di mercato altamente competitiva, con la piena occupazione e il progresso sociale, e la protezione dell'ambiente; a lottare contro l'esclusione sociale e la discriminazioni; a promuovere il progresso scientifico e tecnologico; a rafforzare la coesione economica, sociale e territoriale e la solidarietà tra gli Stati membri nel rispetto della diversità culturale e linguistica che ne contraddistingue la natura.
- Costituzione della Repubblica Italiana: tra i suoi principi fondamentali, l'inviolabilità dei diritti dell'uomo e il riconoscimento dell'uguaglianza formale e sostanziale devono essere garantiti nella ricerca e nell'utilizzo di sistemi IA. La centralità del lavoro, inoltre, impone una riflessione sulle conseguenze economiche e sociali dello sviluppo di tali sistemi. Nell'applicazione dei sistemi

d'arma, sottolinea il rifiuto della guerra come mezzo di risoluzione di controversie internazionali. Il rispetto dei diritti dei cittadini alla libertà personale, alla riservatezza e alla libera manifestazione del pensiero deve guidare l'impiego da parte di soggetti pubblici e privati, dei sistemi IA in grado di minacciare queste libertà fondamentali. Per questi soggetti, va raccomandato il rispetto della dignità umana nell'iniziativa economica. Nell'attività di tutela della sicurezza e dell'ordine pubblico, nonché sulle attività di prevenzione e repressione dei crimini, occorre tanto che l'attribuzione di responsabilità civili e penali per un utilizzo improprio di sistemi IA sia personale, quanto rafforzare il rispetto del principio di legalità e del giusto processo.

2. Principi e valori etici

Sulla base dell'analisi svolta nei paragrafi precedenti, il gruppo ha identificato un insieme di principi e valori etico-sociali, organizzati in tre macro livelli con una stratificazione di raggio crescente, da quello individuale a quello globale.
Le tre macro categorie, ed i relativi principi, sono da intendersi secondo una prospettiva integrata, non di rilevanza (Fig. 1: cerchi concentrici). La scelta di un tale approccio non riflette solamente necessità formali, come la compilazione di un aggregato di enunciati, ma sottende una volontà sostanziale: proporre una visione organica.

2.1. Livello individuale

Il livello individuale si propone di identificare i valori che pertengono alla persona e si incardinano sul fondamento monolitico della dignità umana. Da quest'ultima derivano i diritti civili e il principio della non discriminazione, i quali tutti si esplicano sia nella dimensione materiale che nella dimensione immateriale delle attività umane.

2.1.1. Dignità umana

Per quanto sia estremamente difficile accordarsi sulla sua definizione, il principio della dignità umana è largamente diffuso e comunemente riconosciuto. In quanto tale, esso ricopre un ruolo fondamentale nella Dichiarazione universale dei diritti umani, nella Carta dei diritti fondamentali dell'Unione Europea e nella Costituzione della Repubblica Italiana.
Nel suo senso più fondamentale, per dignità si intende il valore intrinseco che pertiene ad ogni individuo in quanto essere umano—valore che, per rifarsi ad una nota formula kantiana, impone di non trattare mai un altro essere umano solo come un mezzo per i propri scopi, ma anche e sempre come un fine in sé, cioè come un soggetto in grado di determinare sé stesso in maniera autonoma. Il principio della dignità umana costituisce una limitazione del potere di autodeterminazione e di azione del singolo, per cui il valore intrinseco del suo simile funge da confine della propria libertà.
L'IA, data la sua pervasività sociale e l'impatto profondo che si ritiene eserciterà su ogni aspetto della vita, potrà avere effetti rilevanti sul rispetto della dignità umana. Applicazioni in campo industriale, sanitario, educativo, assistenziale e sociale potranno offrire nuovi potenti mezzi per la produzione, il mantenimento o il rafforzamento delle condizioni associate alla vita dignitosa. Tuttavia, le stesse tecnologie che possono essere indicate come mezzi per il rispetto e l'affermazione della dignità umana

potrebbero anche minacciarne l'integrità sia morale che fisica, intersecando i temi bioetici della transumanità. Contrarie alla dignità umana sembrano essere tecnologie che manipolano l'utente - anche a fine di bene - o a cui sono delegate decisioni di grande importanza sociale o esistenziale senza che sia possibile comprenderne le dinamiche. Ancora, la dignità umana è messa significativamente a rischio da tecnologie che non colgono il valore intrinseco di ogni individuo dissolvendo la sua particolarità nella generalità di modelli statistici.
In conclusione, il principio di dignità è ampiamente riconosciuto come un'istanza fondamentale per lo sviluppo e l'uso etico dell'IA.

2.1.2. Libertà e diritti civili

Il principio su cui si basa il modello etico dei diritti civili è la dignità della persona. Il modo più affidabile per assicurare la felicità e la giustizia è l'affermazione del valore dell'essere umano che lo differenziano dagli altri esseri naturali e gli conferiscono anche la propria dignità.
Si può dire che la Dichiarazione Universale dei Diritti Umani è considerata l'origine e il nucleo fondamentale di una costruzione etica come base della soluzione ai conflitti della convivenza umana.
La Dichiarazione universale dei diritti umani costituisce un'etica materiale che stabilisce valori, contiene norme che devono essere rispettate, diritti che devono essere garantiti e libertà che devono essere protette.
Storicamente, i primi diritti che si sono sviluppati sono stati i cosiddetti diritti di prima generazione, i diritti di libertà, che limitano il potere dello Stato, come la libertà di pensiero, di coscienza e di opinione, in risposta contro monarchie assolute e regimi dittatoriali. Successivamente i diritti di seconda generazione hanno riguardato i diritti di uguaglianza ed i diritti politici, che assicurano una parità di condizioni nella partecipazione al potere politico.
La terza generazione, il cui valore fondamentale è la solidarietà, comprende i diritti sociali, il diritto alla sicurezza sociale, il diritto al lavoro.
Come si vede, i diritti umani vanno inquadrati in una prospettiva dinamica: si sono evoluti nel corso dell'esperienza storica, ed è ragionevole ritenere che possano continuare a farlo.
I diritti civili si radicano nella Dichiarazione universale dei diritti dell'uomo che riconosce il diritto di tutte le persone alla libertà (di circolazione, di pensiero, di opinione, di associazione, ecc.), alla giustizia, un livello di vita adeguato, alla salute e al benessere, in particolare alle cure mediche e ai servizi sociali. Tutti questi sono ambiti in cui le tecnologie assumono un ruolo preponderante. Grazie alla IA, la dimensione immateriale è infatti divenuta (o sta divenendo) la principale interfaccia utente per le relazioni sociali ed economiche delle persone, la sede prima in cui tali diritti vanno assicurati (Quintarelli, 2019). Tale assicurazione deve essere sostanziale, prima ancora che formale, bilanciando pertanto gli squilibri esistenti nella dimensione materiale tra diversi individui ed includendo una cautela particolare per le persone più deboli che statisticamente sarebbero relegati ad *outlier* nei modelli statistici.

2.1.3. Non discriminazione

La Carta dei diritti fondamentali dell'Unione Europea afferma il principio di uguaglianza: riconoscere a tutti i cittadini gli stessi diritti davanti alla legge. Il principio della parità tra uomo e donna è alla ba-

se di tutte le politiche continentali, ed è l'elemento su cui si fonda l'integrazione europea. Si applica in tutti i settori.
La Costituzione italiana recita:
«Tutti i cittadini hanno pari dignità sociale e sono eguali davanti alla legge, senza distinzione di sesso, di razza, di lingua, di religione, di opinioni politiche, di condizioni personali e sociali. È compito della Repubblica rimuovere gli ostacoli di ordine economico e sociale, che, limitando di fatto la libertà e l'eguaglianza dei cittadini, impediscono il pieno sviluppo della persona umana e l'effettiva partecipazione di tutti i lavoratori all'organizzazione politica, economica e sociale del Paese».
Ed anche:
«La Repubblica riconosce a tutti i cittadini il diritto al lavoro e promuove le condizioni che rendano effettivo questo diritto».
Da queste previsioni consegue la lotta contro la discriminazione, la tutela dei diritti delle minoranze e dei settori più fragili della popolazione in relazione alla loro situazione oggettiva.
I dati rilevati ed utilizzati nei sistemi di *machine learning* dipingono i tessuti sociali incorporandone i relativi pregiudizi. In assenza di specifiche cautele e previsioni, i modelli statistici prodotti cristallizzano e possono amplificare tali bias.

Capire i bias e come gestirli

I sistemi di IA, inevitabilmente, ereditano dagli esseri umani molti dei bias di cui soffrono. Le modalità in cui questi errori vengono trasmessi possono essere molteplici, e difatti possiamo facilmente identificarne almeno cinque (Hammond, 2016): bias nei dati stessi; bias nati da interazioni; bias di similarità; bias che scaturiscono da obiettivi contrastanti; bias emergenti. Nel primo caso, l'errore è contenuto direttamente nei dati che alimentano il sistema, ed è spesso associato a banche dati incomplete, errate, o mal classificate. Nel caso di "bias da interazioni", gli errori sorgono nel momento in cui il programma interagisce con utenti esterni e apprende attraverso tali interazioni (si pensi al caso del *chatbot* Tay, tristemente noto esempio di un *chatbot* che sviluppò ideologie naziste attraverso scambi di opinioni con utenti su Twitter). I bias di similarità, invece, nascono da sistemi che implementano correttamente le azioni per le quali sono stati programmati, ma che involontariamente in questo modo restringono le possibilità del sistema stesso (un esempio è il *news feed* di Google o Facebook, che mostra agli utenti notizie simili). Per quanto riguarda i bias che derivano da obiettivi contrastanti, esistono situazioni in cui un sistema disegnato per assolvere una specifica funzione in un ambito ben delineato crea conseguenze negative in applicazioni laterali e secondarie. Infine, i bias emergenti sono quelli che rafforzano pregiudizi e comportamenti umani che potrebbero essere opinabili.

2.2. Livello sociale

I sistemi di IA, essendo non deterministici, tendono a generare alcune predizioni errate. Nella valutazione delle loro conseguenze si determina una tensione tra il livello individuale ed il livello sociale.

Questa sezione definisce i valori rilevanti da rispettare e promuovere in un'ottica di bilanciamento tra bene comune e individuale.

2.2.1. Inclusività

Dal punto di vista dell'equità, l'adozione dell'IA deve garantire a tutti, incluso categorie deboli o svantaggiate, accesso equo a opportunità, servizi e del lavoro prodotto, evitare concentrazioni di risorse e potere.
Gli effetti dello sviluppo dell'IA sulle diseguaglianze riguardano infatti non soltanto l'aspetto monetario, ma anche quello relativo a questioni socio-culturali, dove a condizioni di disagio si sovrappongono la mancanza di accesso all'educazione digitale e discriminazioni su base etnica o di genere. Un uso pregiudizievole delle nuove tecnologie rischia quindi di minare la posizione di categorie in posizione già precaria, innescando un circolo vizioso di marginalizzazione e ulteriore aumento delle diseguaglianze.
Un'adozione equa delle tecnologie intelligenti richiede che i relativi vantaggi e le opportunità connesse – finanziarie, educative, giuridiche, sanitarie, assistenziali, e così via – siano inclusive di un numero più ampio possibile di cittadini, a prescindere dalla loro condizione sociale, classe di reddito. ubicazione geografica e da altri fattori analoghi (*Sustainable Development Goals*).

2.2.2. Riduzione disuguaglianza

L'IA dovrebbe essere sviluppata in modo da prevenire e ridurre attivamente le diseguaglianze, garantendo la massima condivisione dei benefici socio-economici delle nuove tecnologie e vigilando affinché i guadagni di produttività garantiti dalla sua implementazione non diventino monopolio di una ristretta cerchia di soggetti, ma siano invece distribuiti equamente attraverso diverse categorie e classi sociali.
L'IA può diventare una forza attiva per la riduzione delle diseguaglianze, incorporando un concetto di giustizia distributiva che guardi alle categorie marginali come soggetti di intervento prioritario: strumenti basati sull'IA, ad esempio, possono risultare utili, nell'ambito del sistema educativo, per colmare divari di apprendimento, mentre in ambito sanitario, possono essere utilizzati per stimolare il *social empowerment* ed i servizi destinati agli individui affetti da disabilità.
L'impatto sul mondo del lavoro, particolarmente rilevante in relazione alla Costituzione Italiana, richiede non solo particolare attenzione affinché siano adeguatamente ammortizzati fenomeni negativi quali esuberi di massa, disoccupazione generalizzata, *de-skilling* e deprezzamento del lavoro umano, ma anche la definizione di nuove *policy* del lavoro che affrontino la relazione tra uomo e macchina nonché l'elaborazione di nuovi dispositivi sociali capaci di mitigare le esternalità negative dell'automazione e favorire generalizzate condizioni di esistenza dignitose.

2.2.3. Coesione sociale

Lo sviluppo dell'IA deve favorire la coesione sociale e garantire la robustezza del procedimento democratico. La ricerca (Sirbu et al., 2019) mostra come il design dei sistemi di IA usati in una comunità possa facilitare la formazione di un consenso in un numero ridotto di interazioni ma sia anche in gra-

do, viceversa, di rafforzare e radicare le divisioni nel tempo. Questo effetto, noto e sfruttato da tempo nei mass media, assume una rilevanza estremamente significativa nell'era dei *personalized* media. In tale contesto, obiettivi socialmente desiderabili e legittimi interessi aziendali possono divergere ed entrare in conflitto. Da una parte, l'interesse sociale nel rafforzamento della coesione sociale suggerisce l'adozione di tecnologie capaci di favorire la composizione di opinioni differenti e promuovere confronti tolleranti. Al contrario, per massimizzare l'*engagement* degli utenti, la quantità di interazioni, e quindi lo *screen time* e i ricavi collegati, le aziende sono portate a adottare tecnologie IA tese ad amplificare le divisioni ed esacerbare gli animi, caratteristiche sfruttate per la diffusione delle cosidette *fake news* e *deep fakes*.
Un'altra modalità in cui si esplica questa divergenza di obiettivi è l'uso di algoritmi in grado di sfruttare il *confirmation bias* degli utenti effettuando una iper-personalizzazione dei messaggi. La riproposizione mirata di contenuti affini (c.d. *echo chamber*), giustificata dall'esigenza di migliorare l'esperienza dell'utente, rischia di compromettere il pluralismo informativo.
Questi fenomeni, se non gestiti, influenzano negativamente i processi democratici e minano la coesione sociale con effetti socioeconomici profondi e di lungo termine.

2.3. Livello globale

In continuità con i principi discussi a livello individuale e sociale, prevenzione del danno, ricerca di pace e giustizia e sostenibilità si configurano infine come cardini globali dello sviluppo etico dell'IA.

2.3.1. Prevenzione del danno

I sistemi informatici consentono di affrontare problemi con una scalabilità sostanzialmente illimitata, ben superiore a quella possibile agli esseri umani. I sistemi di IA consentono di affrontare problemi di natura differente rispetto ai tradizionali domini applicativi dei sistemi algoritmici deterministici, quali ad esempio i problemi di percezione e classificazione, precedentemente riservati alla attività umana, che possono così essere oggetto di una scalabilità sostanzialmente infinita a livello globale. Ciò aumenta le possibilità umane ma ne espande i possibili rischi: così come è globale l'utilizzo delle tecnologie, è globale la propagazione di eventuali errori e dei danni relativi.
La prevenzione del danno si concretizza in una valutazione dei rischi finalizzata a adottare o applicare misure che ne prevengano la manifestazione o ne mitighino l'esposizione o gli effetti. La prevenzione diventa quindi un processo dinamico che periodicamente valuta i sistemi attraverso procedure di *risk assessment*, promuovendo procedure e protocolli per il *risk management*. La creazione di buone pratiche come quelle introdotte nel campo della sicurezza informatica (ISO/IEC 27001) permette di identificare e descrivere situazioni critiche. Le pratiche di *risk management* adottano e promuovono "gli scenari di rischio" come metodologia utile per l'analisi del rischio.
Diventa di attuale interesse raccogliere e prototipare scenari che possano essere utilizzati per valutare i sistemi autonomi in modo tale da poter definire diverse classi di rischio. Tra gli scenari ormai conosciuti che possono essere censiti e raccolti per una periodica valutazione rientrano la trasmissione di bias ai dati utilizzati per il *training*, il *data poisoning*, lo *adversarial attack*.

2.3.2. Pace e giustizia

L'articolo 3 del Trattato Fondamentale dell'Unione Europea sancisce come obiettivi condivisi dagli Stati membri la promozione della pace e della giustizia, in aderenza alla Convenzione Europea sui diritti umani ed alla Carta delle Nazioni Unite.

La rivoluzione tecnologica legata all'IA corre il rischio, da un lato, di accentuare le disuguaglianze interne ai paesi avanzati; dall'altro, di scavare un solco ancora più profondo tra questi ed i paesi in via di sviluppo.

Uno sviluppo etico dell'IA deve quindi garantire la protezione dei valori della concordia e della fratellanza. Inoltre, l'innovazione tecnologica deve essere condotta nel rispetto dei principi di giustizia e contribuire a prevenire l'erompere di conflitti e tensioni internazionali.

Nella prevenzione e nella repressione dei crimini, i principi di legalità e di "giusto processo", garantiti dal diritto internazionale, dalla Costituzione e riconosciuti dagli ordinamenti, devono collocarsi come presupposti irrinunciabili al dispiegamento di sistemi IA in questi settori.

AWS: Autonomous Weapon Systems

L'applicazione di IA e sistemi autonomi promette di rivoluzionare il contesto dei conflitti militari e della *governance* di sicurezza degli attori statali. Un approccio etico in questo contesto deve preservare l'autonomia e la determinazione umana al fine di garantire sempre il controllo dei meccanismi decisionali autonomi e la responsabilità soggettiva, quantomeno in quegli ambiti ritenuti critici dal punto di vista strategico e, soprattutto, morale, rispecchiando l'esigenza di proteggere valori di rilevanza costituzionale quali la vita, l'incolumità fisica e la dignità umana.

La garanzia del controllo umano significativo sui sistemi autonomi è necessaria al fine di preservare la natura benefica della tecnologia, in linea con i principi prevalenti nel campo della bioetica e dell'etica della tecnologia.

Ciò è valido in particolare nel caso delle *Lethal Autonomous Weapon Systems* (LAWS), i sistemi d'arma autonomi. Il divieto di produzione ed utilizzo di LAWS rispetta gli standard prevalenti di *human-on-the-loop* (HOTL) e *human-in-the-loop* (HITL), secondo cui l'applicazione di sistemi autonomi in ambiti *safety-critical* deve avvenire sempre sotto la supervisione ed il controllo di un operatore umano.

In ogni caso, i sistemi autonomi progettati per arrecare un danno fisico ad infrastrutture e persone militari o militarizzate hanno speciali e inconsueti risvolti etici se paragonati con le armi di tipo tradizionale o con i sistemi non armati. Per questo motivo è necessario da un punto di vista etico poter garantire almeno i seguenti requisiti:

a) La finalità difensiva dell'esercizio dei sistemi autonomi (Ovvero la natura di *Defensive Autonomous Weapon Systems*).
b) Assicurare il loro controllo e la validazione finale della decisione esecutiva da parte dell'uomo (***controllo***).
c) Progettarli in modo che abbiano sistemi di tracciamento che garantiscano l'attribuzione delle responsabilità nel loro uso (***responsabilità***).

d) I loro sistemi di apprendimento e adattamento devono documentare ed essere in grado di spiegare in modo comprensibile all'operatore umano le loro determinazioni (***trasparenza e spiegabilità***).
e) Sulla scorta del precedente, si deve fare in modo che l'operatore umano preveda il comportamento delle loro funzioni autonome (***fiducia***), tanto in ambito direzionale quanto in quello operativo.
f) Bisogna sviluppare codici etici professionali ed addestrare operatori umani che siano responsabili del loro uso e siano chiaramente identificabili (***formazione***).

2.3.3. Sostenibilità

Il più grande singolo fattore tecnologico che favorirà il raggiungimento dei 17 obiettivi di sviluppo sostenibile (SDG) nei prossimi anni sarà la rivoluzione digitale, determinata dai continui progressi nel campo dell'informatica ed in particolare del *machine learning* e della robotica. La rivoluzione digitale rivaleggia con il motore a vapore, il motore a combustione interna e l'elettrificazione per gli effetti pervasivi su tutti i settori dell'economia e della società e pertanto impattante molti ambiti indirizzati dai 17 SDG.
Ad esempio, la IA permeerà in misura crescente quasi tutti i settori dell'economia, dall'agricoltura (agricoltura di precisione), all'industria mineraria (veicoli autonomi), alla produzione (robotica), alla commercializzazione (profilazione), alla finanza (modelli comportamentali), ai media (*targeting* individuale), alla salute (diagnostica), ecc.
In generale, questi contributi della tecnologia possono aumentare la produttività, ridurre i costi di produzione, espandere l'accesso a beni e servizi, dematerializzare la produzione riducendo l'impatto sull'ambiente, migliorare il funzionamento dei mercati, migliorare la ricerca e le terapie farmacologiche, semplificare l'accesso ai servizi pubblici, ecc.
Tuttavia, vi sono anche evidenti rischi e svantaggi della rivoluzione che devono essere identificati e affrontati. Forse il più temuto è la perdita di posti di lavoro e lo spostamento della distribuzione del reddito dal lavoro al capitale.
I processi di automazione sono in corso da decenni e una conseguenza importante, a quanto pare, è la riduzione della domanda di lavoratori meno qualificati. Con i progressi nell'IA e nella robotica, molti più lavoratori possono ora vedere minacciati il loro lavoro e i loro redditi.
Mentre i nuovi posti di lavoro potrebbero sostituire quelli vecchi, i nuovi posti di lavoro potrebbero avere redditi reali e condizioni di lavoro più bassi.
Ci sono molte altre minacce percepite dalla rivoluzione digitale. Le identità digitali possono essere rubate. I governi e le imprese private possono invadere la privacy e monitorare gli individui contro la loro volontà o a loro insaputa. Alcune aziende possono sfruttare i loro vantaggi nell'accumulare grandi dati per conquistare una posizione dominante di monopolio nei rispettivi mercati, consentendo loro di porsi sostanzialmente al riparo rispetto alla concorrenza da parte di nuovi entranti inficiando il funzionamento del mercato.
I *social media* possono essere manipolati e *cyber* attacchi possono paralizzare una società interrompendo i flussi di informazioni o colpendo i dispositivi collegati a Internet.

Il problema dell'impatto ambientale delle tecnologie di IA richiede particolare attenzione in ragione dell'elevato consumo energetico.
Per i progetti che coinvolgono Paesi meno avanzati o in via di sviluppo, è fondamentale tenere in considerazione se e in che modo i sistemi IA possano integrarsi con le soluzioni già adottate in questi contesti e quali risorse siano necessarie per la loro effettiva implementazione. In particolare, la scarsa quantità e qualità delle informazioni digitali raccolte presso le zone di intervento può ostacolare l'adozione di sistemi IA. Occorre riflettere sull'eventualità di mettere a disposizione dei Paesi meno avanzati o in via di sviluppo dati semanticamente interoperabili raccolti o elaborati da Stati tecnologicamente più avanzati.

3. Diritti

In questa sezione si elencano i diritti che discendono dai principi e valori etici precedentemente analizzati e che devono informare le raccomandazioni circa lo sviluppo etico delle tecnologie IA.

3.1. Informazione

Tutti i sistemi autonomi prevedono che l'individuo abbia uno scambio attivo di informazioni, utili ad elaborare lo scenario e suggerire una soluzione al problema affrontato.
È un diritto dell'utenza conoscere ed essere informato sull'intero processo: dalla raccolta dei dati e delle informazioni, dal procedimento di elaborazione ai rischi, e sulla stessa natura dell'interazione con il sistema (dove questa sia con un sistema autonomo in grado di elaborare le informazioni o meno). Il consenso informato dovrebbe essere presentato in modo chiaro e succinto permettendo una scelta consapevole ed evitando adesioni irriflesse.
Nel caso di decisioni che possono impattare significativamente sulla vita degli utenti o sulla società nel suo complesso, è necessario tutelare il diritto di scelta circa il livello di autonomia/intelligenza del sistema durante l'interazione, specificando le conseguenze della scelta, nonché di poter richiedere un intervento totalmente umano.

3.2. Educazione

La consapevolezza riguardante potenzialità e rischi legati alla tecnologia va di pari passo con l'educazione e la formazione tecnologica.
È desiderabile educare, istruire e formare la società e le persone ad un uso corretto e ad una coesistenza matura con la tecnologia. Differenziare educazione, istruzione e formazione permette di considerare separatamente aspetti importanti del rapporto uomo-macchina.
In questo contesto, educare significa saper inquadrare i rapporti tra persone e tecnologia, in particolare come l'individuo dovrebbe rapportarsi e interagire in modo consapevole con gli strumenti forniti. Istruire significa saper trasmettere i saperi che permettono alla persona di conoscere (anche in modo sommario o generale) come funziona la tecnologia e di conseguenza di valutarne rischi e potenzialità. Formare si riferisce ad un processo di apprendimento attraverso il quale l'utenza (conscia delle proprie conoscenze e lacune) migliora e incrementa la propria istruzione.

In quest'ottica l'abuso della tecnologia diventa quindi una carenza di educazione, mentre il proliferare di allarmismi catastrofici una mancanza di istruzione della tecnologia.
Per questo, alzare il livello di *information literacy* risulterebbe in una maggiore adeguatezza e consapevolezza degli individui rendendoli maggiormente adeguati ad affrontare la rapidità di evoluzione del mondo.

3.3. Autodeterminazione dell'identità

Proponiamo di usare come definizione di identità di una persona l'insieme di attributi materiali ed immateriali che la definiscono descrivendone unicità e diversità. L'autodeterminazione dell'identità è così un diritto imprescindibile e inalienabile radicato nel diritto alla dignità di ciascun individuo.
La natura sociale del vivere umano rende l'identità personale una caratteristica anche sociale, ovvero plasmata attraverso un percorso dove libere interazioni con l'ambiente e altri individui permettono di costruire una narrativa che cambia con il tempo e l'ambiente stesso.
La raccolta di dati e l'interazione con la tecnologia nelle pratiche quotidiane rende l'intelligenza artificiale uno strumento potente in grado di svolgere mansioni e soddisfare bisogni migliorando la qualità della vita, ma allo stesso tempo può divenire un mezzo per manipolare le decisioni degli individui minandone l'autodeterminazione (Taddeo/Floridi, 2018). Appartengono, inoltre, al medesimo ambito aspetti più pragmatici della tutela dell'identità quali il diritto alla portabilità, alla rettifica, all'oblio ed altri diritti connessi, giuridicamente tutelati.
Per rendere azionabili queste forme di tutela, l'individuo deve essere sempre messo in condizione di sapere e conoscere la natura del suo interlocutore (artificiale o meno), le sue finalità e potenzialità, al fine di consentirgli di scegliere come interagire con l'agente e quali facoltà accordargli sui propri dati.

3.4. Riservatezza

La tutela della riservatezza si declina nel riconoscimento di una sfera privata all'interno della quale l'individuo deve essere immune da intromissioni di terze parti, siano esse pubbliche o private.
L'aumento della capacità di calcolo, di *storage* e di connessione determinato dall'evoluzione dell'elettronica porta all'accumulo di dati nel tempo, provenienti da dispositivi, sensori e sonde di ogni tipo, che si fondono con l'ambiente materiale in cui le persone vivono e l'ambiente immateriale in cui esse operano; l'aumento della capacità di elaborazione, grazie ai modelli statistici che ne vengono distillati, permette di passare da applicazioni algoritmiche deterministiche ad applicazioni probabilistiche.
Ci sono quindi tre dimensioni la cui rilevanza cresce esponenzialmente rispetto alle applicazioni informatiche tradizionali: la compenetrazione spaziale, l'accumulo temporale e la modellazione statistica.
La prima dimensione impone una re-ingegnerizzazione concettuale dello spazio all'interno del quale l'individuo si muove. Con la progressiva erosione della barriera che separa immateriale e materiale, occorre valutare l'impatto dei sistemi IA nella sfera privata dell'individuo considerando l'interconnessione tra l'ambiente fisico in cui hanno luogo i suoi movimenti e la percezione digitalizzata che di essi hanno i sistemi stessi.

Una seconda prospettiva ruota intorno al tempo come elemento distintivo dell'analisi. L'utilizzo di dati riferiti al passato per lo sviluppo automatico di modelli predittivi richiede una riflessione sull'effetto di cristallizzazione sociale che si potrebbe determinare se le valutazioni più invasive della riservatezza fossero prese senza analizzare la correzione, l'integrazione o l'eliminazione di bias o dati non più rilevanti. Nei casi di accumulo permanente e indiscriminato di dati (*always on*), occorre riflettere quali concrete possibilità di sottrarsi, anche temporaneamente, alla raccolta di informazioni siano da garantire agli individui per ragioni di riservatezza.
Una terza prospettiva è legata ai profili cognitivi dell'IA e all'estrazione di nuova conoscenza a partire dai dati. Nella fase di valutazione della pervasività dei sistemi IA nella sfera privata dell'individuo, occorre porre l'accento non solo sulla gestione del dato personale osservato, ma anche sull'impatto delle inferenze che tali sistemi sono in grado di generare e sull'ottica super-individuale attraverso cui essi consentono di osservare la realtà.

Anonimato protetto (*differential privacy*)
Sempre più spesso emerge la delicatezza dell'equilibrio tra la necessità di poter perseguire alcuni reati e la garanzia della libertà di espressione rispetto al rischio di censura. Il concetto di reputazione, ovvero di fiducia attribuita ad un soggetto, è uno dei pilastri centrali della società in tutti i suoi aspetti relazionali, dall'economia alla politica.
La Costituzione italiana recita:
«Tutti hanno diritto di manifestare liberamente il proprio pensiero con la parola, lo scritto e ogni altro mezzo di diffusione».
Ed anche:
«La responsabilità penale è personale».
Rispetto al diritto alla libertà di espressione, le valutazioni circa la rimozione di contenuti sono affidate a moderatori che operano globalmente secondo principi definiti dalle piattaforme, non necessariamente allineati con le disposizioni dei governi locali, e con quanto stabilito dagli ordinamenti.
La libertà di espressione non deve esporre a possibili ritorsioni personali da parte di gruppi di pressione. Nel contempo, deve essere altresì garantita la possibilità, nel caso in cui il fatto costituisca reato, che l'autorità giudiziaria possa risalire alla identità di chi lo ha commesso.
Queste esigenze sono contrastanti e difficilmente conciliabili in quanto gli attuali servizi digitali e l'infrastruttura giuridica non sono state disegnate in modo coerente per gestirle. Il rapporto tra anonimato e conoscibilità è oggetto è riferibile al paradigma di *differential privacy* o anonimato protetto.

3.5. Tutela dei diritti

Una efficace tutela dei propri diritti è esso stesso un diritto fondamentale delle persone, in linea con l'obiettivo 16 degli SDG e la necessità di assicurare effettività alle condizioni essenziali della democrazia (Rodotà, 2012, pp. 63).

Lo sviluppo dell'elettronica determina un aumento delle capacità di calcolo, di archiviazione e di comunicazione dei dispositivi elettronici ovvero una riduzione del loro costo rendendoli pervasivi ed in grado di catturare dati da ogni azione ed interazione. Dal computer su ogni scrivania, si passa ad un computer pervasivo in cui le funzioni ed i dati delle persone sono distribuiti e l'accesso a tali funzioni e dati è assicurato tramite interfacce uomo-macchina semplificate e basato sul riconoscimento dell'identità.

Grazie a questa disponibilità di dati e capacità di calcolo, le tecniche di IA consentono di realizzare software non più algoritmicamente deterministici ma basati sull'applicazione di modelli statistici distillati dai dati. La natura delle applicazioni realizzabili cambia, consentendo di applicare la scalabilità tipica delle applicazioni informatiche ad ambiti differenti non indirizzabili con precedenti algoritmi deterministici, tra cui, ad esempio, casi tradizionalmente vincolati alla percezione umana e la loro classificazione.

In questo genere di usi dell'IA, il fattore di scala muta la natura stessa dell'applicazione. Si consideri ad esempio il passaggio dell'esame delle foto segnaletiche, effettuato da persone e quindi limitato a poche decine di migliaia di individui, alla sua effettuazione con strumenti IA, indefinitamente scalabile, e quindi potenzialmente riguardante molti milioni di persone. Nel definire le riserve di legge, i padri costituenti potevano prevedere esclusivamente l'intervento umano che implicitamente incorpora frizioni e limiti alla scalabilità. Con la scalabilità offerta dall'IA la natura del controllo può mutare da eccezione a regola sociale che non può essere considerata alla stregua di un semplice *upgrade* tecnologico ma che pone nuovi interrogativi in merito alla tutela dei diritti delle persone.

3.6. Diritti dei soggetti deboli

Alcune categorie di soggetti, quali minori, anziani, e non autosufficienti si possono trovare nell'incapacità di prendere decisioni autonome, incapacità legata all'età o a condizioni psico-fisiche.

La Convenzione ONU sui diritti dell'infanzia e dell'adolescenza stabilisce in particolare come l'individuo di minore età sia titolare di diritti imprescindibili oltre che un elemento al quale garantire particolare tutela in quanto oggetto di cure e assistenza speciali [11]. L'art. 29 sancisce l'importanza dello sviluppo della personalità, identità e attitudini con speciali riferimenti all'ambiente e alle relazioni interpersonali descritte attraverso diversi livelli di interscambio culturale.

Oggi si osserva come, generalmente, la tecnologia, ed in particolare l'IA, sia invece progettata e costruita per catturare l'attenzione (soprattutto dei più giovani) e mantenere quest'ultima il più a lungo possibile al fine di raccogliere o produrre più materiale informazionale possibile.

La tecnologia dovrebbe invece considerare queste categorie un elemento di forte tutela, promuovendo valori e metodologie che lo inquadrino in un ambiente di assistenza e crescita, favorendone lo sviluppo cognitivo e che ne permetta la libera determinazione, senza vincolarne scelte, preferenze ed attitudini, in particolare se strumentali ad uno sfruttamento commerciale o con finalità manipolative (SDG 3).

4. Obblighi e raccomandazioni

Alla luce dei principi e dei diritti identificati in precedenza, questa sezione individua obblighi e propone raccomandazioni a cui attenersi nello sviluppo etico e nell'utilizzo regolamentato di sistemi di IA.

4.1. Fiducia

La fiducia gioca un ruolo cruciale in ogni processo di innovazione: solo tutelando e promuovendo il capitale di fiducia sociale relativo all'IA sarà possibile coglierne appieno le potenzialità. Di conseguenza, è necessario individuare i fattori che potrebbero minare la fiducia in questa tecnologia e mettere in campo misure efficaci che ne limitino o eliminino le esternalità negative.
La fiducia è un collante sociale primario e sostiene l'organizzazione del lavoro, la divisione dei compiti e la delegazione delle mansioni, rendendo disponibili a livello comunitario opportunità e prospettive altrimenti irraggiungibili.
Data la rilevanza della fiducia in ogni interazione umana, è necessario prendere precauzioni affinché anche le tecnologie basate sull'IA — in quanto mediatori di relazioni sociali — risultino affidabili, degne della fiducia dei diversi attori coinvolti nel loro utilizzo (SDG 9).
Le tecnologie realizzate devono rispecchiare i valori degli utenti e della società nel cui contesto vengono utilizzate, conformemente a obiettivi socialmente condivisi.
È necessario che siano definiti *framework* di fiducia, ovvero corpus di metodologie, regole, certificazioni, controlli, sanzioni, *benchmark* finalizzati al raggiungimento di target socialmente desiderabili e politicamente determinati.

4.2. Accessibilità

Essendo assai rilevante l'impatto sociale delle decisioni prese per mezzo di algoritmi, è fonte di grande preoccupazione il fatto che le modalità tramite cui tecniche di apprendimento automatico processano informazioni non siano facilmente accessibili, ovvero non siano trasparenti e spiegabili. Questa è la ragione per cui si raccomanda a più voci che i sistemi decisionali autonomi siano i più trasparenti possibile e che le loro determinazioni possano essere spiegabili.
L'opacità dei modi in cui questi sistemi elaborano i dati su cui sono basate le decisioni prodotte — che fa di essi delle *black box* (Pasquale, 2015) — è problematica sia da un punto di vista tecnico che da un punto di vista sociale: l'assenza di accessibilità può ingenerare il sospetto che alcune correlazioni su cui la decisione è basata possano incorporare pregiudizi eticamente condannabili e causare discriminazioni, trattamenti non equi e ingiustizie.
Sul piano tecnico è necessaria la trasparenza per rendere conoscibile come funzioni il processo decisionale del sistema, la sua logica interna, e per poterlo validare da un punto di vista tecnologico.
Sul piano sociale la spiegabilità attiene alla traduzione della funzione dell'algoritmo in termini comprensibili agli utenti, per poter fornire le motivazioni dell'output.
Infatti, si può avere spiegabilità senza trasparenza ad esempio quando una decisione venga presa sulla base di un criterio noto ma senza accesso all'algoritmo che lo elabora.
Si può avere trasparenza senza spiegabilità laddove venga fornito un pieno accesso ai dati ed agli algoritmi ma la determinazione dell'output non sia motivabile dal sistema. Il valore della trasparenza

richiede lo sviluppo di nuove tecniche di spiegazione capaci di aprire le black box e rendere conto dei loro processi interni.
La garanzia dell'accessibilità è inscindibile dall'elaborazione di policy che, includendo trasparenza e spiegabilità, rafforzi la fiducia e protegga il diritto degli utenti – cittadini, autorità e comunità scientifica – ad essere informati in modo semplice e chiaro circa l'uso di intelligenze artificiali e dei limiti collegati a tale uso.

4.3. Sicurezza

La tutela della sicurezza della persona deriva direttamente dai valori fondamentali della dignità personale e come tale richiede di essere rispettata in ogni fase del processo di design e utilizzo delle tecnologie IA. Di conseguenza, concentriamo l'attenzione sulla sicurezza intesa come integrità del sistema.
La pervasività della tecnologia rende la sicurezza un obbligo imprescindibile che il fornitore, di servizi o della tecnologia, deve assicurare a diversi livelli di applicabilità: dalla tutela della sicurezza degli individui alla conservazione dei dati personali, dalla protezione e gestione degli asset fisici all'integrità strutturale del sistema.
Per quanto riguarda i dati personali, ricordiamo le indicazioni riportate dal Regolamento Ue 2016/679 [7]. Tra gli altri diritti garantiti dal testo normativo, è necessario predisporre misure idonee a limitare il trattamento di dati personali. Tale facoltà assume un particolare significato di tutela della sicurezza dei dati nei confronti del loro impiego da parte di sistemi informatici (Rodotà, 2012, pp. 399) e, a maggior ragione, di sistemi di IA.
Un sistema sicuro dovrebbe poi escludere ed evitare possibili esternalità negative, come anche essere strutturato in modo tale da non incentivare il sistema stesso o terze parti a raggiungere l'obiettivo prefissato con strumenti o azioni non idonee. Allo stesso tempo, dovrebbe essere consentito al sistema di apprendere nuove strategie per completare un'azione senza che queste abbiano ripercussioni inattese, e garantirgli un grado di flessibilità tale da potersi adattare a diverse situazioni senza dover essere controllato ad ogni passo, specie qualora i meccanismi di monitoraggio siano estremamente complicati o dispendiosi.
Identificare accuratamente il livello di sicurezza di un sistema è di capitale importanza tanto quanto fare in modo di comunicare tale livello agli utenti in modo intuitivo e chiaro. Da questo punto di vista, dovrebbe essere elaborato un linguaggio funzionale – ad esempio, tramite il ricorso a certificazioni o etichette – che semplifichi la comprensione del livello di sicurezza ed affidabilità di una tecnologia IA.

4.4. Usabilità

È risaputo che alcuni sistemi IA si fondano su complessi modelli computazionali, talvolta poco interpretabili e opachi (*black box*), che possono ingenerare una percezione di mancato controllo nella fase di esecuzione.
Se, da un lato, la rilevanza dell'intermediazione tra sistema ed essere umano trova una sua giustificazione nell'esigenza di assicurare trasparenza, controllo e trust, dall'altro essa acquista ancor maggiore significato quando speciali categorie di utilizzatori usufruiscono del sistema. I benefici offerti dai sistemi IA devono essere accessibili a persone non autosufficienti o portatori di handicap, per garan-

tire la massima espressione del loro potenziale e una vita significativa (art. 3 Cost., Art. 26 Carta dei Diritti Fondamentali dell'UE). Speciale attenzione, inoltre, va posta sull'utilizzo consapevole di sistemi IA da parte dei minori.
Occorre porre in bilanciamento, da una parte, l'esigenza di adattare i sistemi alle diverse fasi di sviluppo del minore (ad esempio, sviluppo del linguaggio o del pensiero matematico) e alle sue condizioni soggettive (diversità culturali e di linguaggio o disturbi di apprendimento) e, dall'altra, i limiti da porre allo sviluppo di interfacce manipolative o in grado di ingenerare confusione circa la natura artificiale del sistema (SDG 4).
Risulta quindi necessario, per una corretta interazione tra umani e sistemi di IA, che le interfacce pongano il fruitore umano del sistema al centro del progetto di sviluppo.

4.5. Controllo

Una partecipazione attiva di un essere umano nelle decisioni prese da sistemi IA è necessaria affinché l'operatore non rivesta il ruolo di esecutore passivo esente da responsabilità morali e giuridiche connesse all'utilizzo di tali sistemi. Tale supervisione previene la riduzione dei destinatari delle decisioni a mere variabili di un calcolo probabilistico, condizione inaccettabile nel caso di situazioni critiche in cui è presente un rischio per valori etici di rilevanza costituzionale o in cui sia necessaria una valutazione morale delle conseguenze della decisione. La discussione sul controllo dei sistemi IA può articolarsi in due direzioni: una descrittiva, in cui si definisce il grado di autonomia di un sistema IA (ad esempio, tramite una grandezza ordinale in cui a ciascun valore corrisponde un certo grado di autodeterminazione della macchina e, per converso, il grado di controllo umano (SAE International, 2018)); una normativa, in cui si trasla ogni livello di controllo in un determinato regime giuridico attribuendo responsabilità differenziate e appropriate al contesto di utilizzo.

4.6. Responsabilità

Il dilemma della responsabilità è sicuramente uno degli aspetti più problematici nello sviluppo di nuovi sistemi di IA. Non è tuttavia chiaro se la responsabilità per alcune decisioni prese da un sistema intelligente debba essere attribuita al suo sviluppatore, al venditore del software, all'utilizzatore o a terze parti. Inoltre, è interessante notare come diverse persone abbiano sviluppato una naturale e irrazionale "avversione verso gli algoritmi", che introduce un elemento aggiuntivo di responsabilità: se un medico decide di non seguire la raccomandazione di un sistema di IA non reputato affidabile, ma sbaglia, può essere ritenuto responsabile per gli esiti di tale decisione? Fino a che punto, quindi, possiamo ignorare un sistema di IA? É chiaro che il problema non sia di facile risoluzione, e che probabilmente sia difficile se non controproducente creare e utilizzare un unico sistema di riferimento per gestire la responsabilità di attori diversi in circostanze differenti. Potrebbero essere richiesti, infatti, molteplici framework che considerino contemporaneamente non solo la responsabilità come tale, ma anche la trasparenza, equità, accuratezza e il grado di controllo di un algoritmo.
Per le applicazioni che possono avere un impatto significativo su società, persone e cose occorre attribuire *ex ante* le responsabilità connesse all'impiego di tali sistemi invece di attendere la valutazione *ex post* di un soggetto chiamato ad allocare responsabilità oggettive. I termini contrattuali che

dettagliano diritti, facoltà, immunità e privilegi devono essere chiari ed accessibili, soprattutto nelle applicazioni che sono destinate o che incidono su un ampio numero di persone.
Inoltre, è necessario elaborare meccanismi di *accountability* che impediscano strategie di deresponsabilizzazione o di attribuzione di responsabilità a soggetti non umani.

4.7. Riparazione

Come abbiamo il principio della *privacy by design* per i sistemi di gestione dei dati personali, è opportuno considerare, per i sistemi basati sull'IA che prendono decisioni che possono influenzare la vita delle persone, l'introduzione di meccanismi di riparazione (*redress*) sulla base di un principio di *redress by design*.
La considerazione di base è che un sistema di IA non difettoso, perfettamente funzionante, effettuerà predizioni errate.
Ciò accade, sia a causa del bias presente nei dati di addestramento, sia perché un sistema di questo tipo è inerentemente non deterministico come, per esempio, un autovelox può essere: se si supera il limite di velocità con la propria auto, l'autovelox lo rileva e si ottiene una multa. È possibile fare ricorso, ma l'automobilista è ritenuto colpevole fino a prova di innocenza perché un sistema deterministico, non difettoso (correttamente configurato, certificato e controllato) stabilisce la sua colpevolezza.
Ma con un sistema di intelligenza artificiale non difettoso, perfettamente funzionante, essendo un motore statistico che produce necessariamente risultati probabilistici, questa decisione potrebbe essere giusta nel 98% delle volte e sbagliata nel 2% delle volte (sarebbe inappropriato classificare queste predizioni errate come sbagli), il che significa che in questo 2%, la persona viene riconosciuta colpevole anche quando non lo è (o non può ottenere un servizio, anche se ha pieno diritto di ottenerlo).
Per la persona, la decisione sbagliata può generare ricadute, superando la portata della decisione stessa, ad esempio generando biasimo sociale, feedback negativi online e altre conseguenze che possono diffondersi nella Rete e diventare impossibili da rimuovere.
Questo 2% di errore tollerato non é da intendersi negativamente, poiché garantisce una flessibilità strutturale all'algoritmo e la capacità di adattarsi e includere nuovi elementi emergenti.
In questi casi errati (possono essere falsi positivi o falsi negativi), la procedura di ricorso può non esistere o, se esiste, può essere inefficace, il suo costo può essere eccessivo, può risultare concretamente non accessibile a tutti, può richiedere un tempo eccessivo o può non rettificare le suddette ricadute.
Una più efficace tutela dei diritti dovrebbe prevedere il principio di *redress by design*, ovvero la previsione, fin dalla fase di progettazione, di meccanismi atti a garantire la ridondanza, sistemi alternativi, procedure alternative, ombudsman, ecc. per poter individuare efficacemente, verificare, correggere le decisioni sbagliate prese da un sistema non difettoso, perfettamente funzionante ed eventualmente, affinare le capacità predittive del sistema.

ISSN 2284-4503

Un esempio di *redress by design*

Le norme di prossima vigenza in materia di *enforcement* del copyright implicano la predisposizione di filtri dei contenuti caricati dagli utenti sulle piattaforme di condivisione. Tali filtri sfruttano l'IA per confrontare i video caricati con un database di *firme* di materiale protetto da copyright. La presenza di sanzioni elevate spingerà i gestori delle piattaforme a configurare i loro sistemi massimizzando il recupero, a scapito della precisione. Video di persone innocenti verranno ritenuti in violazione del copyright a causa di predizioni errate legate all'errore statistico inerente al modello, intaccando la libertà di manifestazione del pensiero degli individui i cui contenuti saranno erroneamente censurati. Sebbene ciò possa essere ritenuto un danno marginale quando esaminato a livello sociale complessivo, per la singola persona sarà una violazione totalizzante di un proprio diritto. Con questo meccanismo si istituisce un primo grado di giudizio in cui, un sistema non deterministico, che è noto commettere errori, stabilisce la illiceità di un comportamento di una persona, fino a prova contraria. La procedura di ricorso avverso le decisioni prevede l'intervento di una autorità nazionale, come ad esempio AGCOM. È prevedibile che la procedura di ricorso sarà effettivamente non accessibile a tutti, non sufficientemente tempestiva o valutata in molti casi eccessivamente complessa.

Un'applicazione del principio di *redress by design* potrebbe prevedere una procedura alternativa di trattamento dei casi in cui il sistema IA predice la violazione: il contenuto potrebbe essere immediatamente pubblicato qualora l'utente accetti di associare la propria identità al contenuto stesso (ad esempio in modo analogo a quanto avviene per gli accessi al Wi-Fi con credenziali via SMS) assumendosi quindi la responsabilità di eventuali illeciti che, con riti e garanzie abituali, verrebbero stabiliti da una corte in grado così di conoscere l'identità del presunto colpevole.

4.8. Titolarità dei dati

I dati pertengono all'individuo che li genera. Diverse sono le realtà che utilizzano il dato al fine di estrarne informazione. L'informazione migliora il livello dei sevizi ottenuti in termini di accuratezza ed affidabilità. Basti pensare per esempio a come, nei moderni sistemi di navigazione, le informazioni aggregate permettano di conoscere lo stato del traffico in un determinato tratto stradale informando gli utenti e abilitandoli ad un eventuale cambio di tragitto, migliorando in questo senso il livello di servizio erogato.

Tale informazione è fondamentale per la definizione di meccanismi in grado di fornire servizi utili alle persone, al fine di migliorarne la qualità di vita. La tutela del dato grezzo, della sua titolarità, del mantenimento e possesso è solo un primo passo non ancora sufficiente a garantire la tutela dell'individuo e della collettività. Con l'aumentare dei dati raccolti ed elaborati aumenta la necessità di definire dei meccanismi tecnici, contrattuali e regolamentari che disciplinino l'estrazione e la gestione dell'informazione, identificandone livelli di accessibilità, modalità di impiego e divulgazione.

Anche data la natura non rivale e solo parzialmente escludibile dei dati, agli utenti devono essere tecnicamente forniti piena trasparenza e controllo dei dati raccolti ed elaborati da sistemi di IA, garanzie che devono essere assicurate a livello contrattuale. Lo scenario può radicalmente mutare con

l'introduzione di sistemi crittografici o sistemi di calcolo distribuiti (overlay network crittografici, crittografia omomorfica o sistemi di IA distribuita). Vanno perciò seguiti con attenzione i loro sviluppi, al fine di consentire agli utenti una piena titolarità in merito alla disponibilità dei propri dati.

4.9. *Governance*

La questione riguardante la governance dell'IA non riflette, tecnicamente, un singolo problema, bensì una moltitudine di aspetti differenti. É infatti una problematica omnicomprensiva che tocca tematiche legate tanto alla giustizia, alla responsabilità, dalle strategie e politiche nazionali che concernono l'IA fino alla sorveglianza "intelligente".

- La considerazione che l'IA determina una scalabilità che trascende i naturali limiti alle attività di percezione e classificazione umane, ponendo pressioni su processi sociali consolidati, suggerisce la opportunità di considerare di disporre l'introduzione di frizioni per limitare o rallentare tale scalabilità nei casi in cui ciò possa determinare esternalità negative (si pensi al sopracitato esempio relativo alle foto segnaletiche o alla diffusione di *fake news* che minano coesione sociale e processi democratici).
- Dalla natura intrinsecamente statistica dell'IA consegue che singole istanze di un sistema non difettoso possano determinare eventi avversi a causa di predizioni errate, nel contempo determinando effetti benefici complessivi assai maggiori rispetto alla situazione precedente. Istanze di un sistema di ausilio alla guida possono errare causando qualche incidente, anche fatale, ma nel complesso del suo utilizzo riducendo grandemente il numero di incidenti e vittime. Per certi versi si tratta di una situazione analoga a quella dei prodotti farmaceutici.
- La governance dell'IA dovrebbe assicurare un'adeguata individuazione, misurazione e classificazione delle previsioni errate causate da sistemi non difettosi, per garantire che rientrino nei valori obiettivo socialmente desiderabili stabiliti grazie a processi democratici (SDG 16). In determinati casi potrebbe essere necessario condurre test di validazione dei sistemi e di misura dei loro effetti prima della loro commercializzazione. In taluni altri dovrebbero essere definite procedure di validazione e valutazione di conformità, almeno per i sistemi che possono avere rischi di impatti significativi su cose, persone e società, consentendo di escludere o attribuire correttamente dolo e negligenza.

Per affrontare tali situazioni andrebbe considerata la realizzazione di una autorità o agenzia incaricata di monitorare la diffusione dell'IA e rilevare le sfide emergenti, fornendo informazioni ai decisori politici ed applicandone le determinazioni. Dovrebbe assicurare il soddisfacimento degli obiettivi a livello di sistema fissati dai responsabili politici in relazione alle classi di applicazioni di IA, sulla base degli effetti sugli individui e sulle organizzazioni sociali. A tal fine potrebbe emettere linee guida e raccomandazioni per gli sviluppatori di sistemi IA ed assistere le aziende nell'applicazione di un approccio basato su valutazioni di rischio ed impatto.

Tale organismo potrebbe infine assicurare un coordinamento con l'Unione Europea ed altri organismi internazionali di standardizzazione. Inoltre, potrebbe giovarsi della collaborazione dei corpi intermedi quali sindacati ed associazioni dei consumatori.

4.10. Formazione

L'IA detiene una notevole rilevanza anche nel settore dell'educazione (SDG 4). Si rende necessario, pertanto, identificare in che misura essa generi opportunità e rischi in settori attinenti.
Il corpo sociale tutto deve poter accedere a percorsi formativi per qualificarsi e riqualificarsi, in modo che l'impatto dell'IA sul mondo del lavoro possa incontrare professionisti preparati a coglierne le opportunità e all'altezza delle sfide che tale tecnologia pone a livello tanto etico quanto sociale, evitando così la formazione di gruppi marginalizzati e incapaci di trovare il proprio ruolo nel nuovo panorama lavorativo.
La formazione degli utenti deve permettere un uso consapevole della tecnologia che non ne demonizzi la natura ma ne evidenzi le potenzialità responsabilizzando l'individuo di fronte a rischi e pericoli.
Il mondo aziendale deve promuovere percorsi educativi volti a facilitare l'integrazione di considerazioni di carattere etico legate alle tecnologie IA in via di sviluppo, sia in sede di design – sostenendo percorsi interdisciplinari e critici – sia in tutti gli altri momenti relativi alla presentazione e alla pubblicizzazione del prodotto.
In ultimo, i *decision maker* di ogni livello devono acquisire piena consapevolezza della natura e del funzionamento dei sistemi di IA al fine di redigere regole adeguate al loro utilizzo.

Bibliografia e Sitografia

Y. BONNET, B. RONDEPIERRE, C. VILLANI, *For a meaningful artificial intelligence: towards a French and European strategy*, 2018, consultabile al link: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

European Commission for the Efficiency of Justice (CEPEJ), *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*, 2018, consultabile al link: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

European Parliament and European Council, *Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data* [repealing Directive 95/46/EC [2016] OJ L 119 (GDPR)], 27 Aprile 2016

Future of Life Institute, *Asilomar AI Principles*, 2017, consultabile al link: https://futureoflife.org/ai-principles/

K. HAMMOND, *5 unexpected sources of bias in artificial intelligence*, 2016, consultabile al link:. https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/

House Of Lords, *AI in the UK: ready, willing and able?*, 2018, consultabile al link: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

IEEE, *Ethically Aligned Design*, 2019, consultabile al link: https://ethicsinaction.ieee.org

Independent High-Level Expert Group on AI set up by the European Commission, *Ethics guidelines for trustworthy AI*, 2019, consultabile al link: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

ISO/IEC 27000 family, *Information security management systems*, consultabile al link: https://www.iso.org/isoiec-27001-information-security.html

Partnership on AI, *Tenets – The Partnership on AI,* 2016, consultabile al link: https://www.partnershiponai.org/tenets/

SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, 2018, consultabile al link: https://www.sae.org/standards/content/j3016_201806/

M. Beltrametti, Chatila R., Chazerand P., Cowls J., Dignum V., M.L. Floridi, Schafer B., *AI4People—An ethical framework for a good AIsociety: Opportunities, risks, principles, and recommendations. Minds and Machines*, 2018, 28, 4, pp. 689-707

L. Floridi, M. Taddeo, *How AI can be a force for good*, in *Science*, 2018, 361, 6404, pp. 751-752

F. Giannotti, J. Kertész, D. Pedreschi, A. Sîrbu, *Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model*, in *PLOS ONE*, 2019, 14, 3, p.e 0213246

F. Pasquale, *The black box society*, Harvard University Press, 2015

S. Quintarelli, *Capitalismo immateriale*, Bollati e Boringhieri, 2019

S. Rodotà, *Il diritto di avere diritti,* Gius. Laterza & Figli Spa, 2015