GRUPPO INDIPENDENTE

DI ESPERTI AD ALTO LIVELLO

SULL'INTELLIGENZA ARTIFICIALE

ISTITUITO DALLA COMMISSIONE EUROPEA NEL GIUGNO 2018

ORIENTAMENTI ETICI PER UN'IA AFFIDABILE

ORIENTAMENTI ETICI PER UN'IA AFFIDABILE

Gruppo di esperti ad alto livello sull'intelligenza artificiale

Il presente documento è stato redatto dal gruppo di esperti ad alto livello sull'intelligenza artificiale (IA). Sebbene i membri di tale gruppo, menzionati nel presente documento, sostengano il quadro di riferimento generale per un'IA affidabile illustrato nei presenti orientamenti, non necessariamente essi condividono ogni singola affermazione contenuta nel documento stesso. .

La lista di controllo per la valutazione dell'affidabilità dell'IA presentata nel capitolo III del presente documento sarà sottoposta ai portatori di interessi affinché la sperimentino (fase pilota) e forniscano un riscontro pratico. All'inizio del 2020 sarà presentata alla Commissione europea una versione di tale lista di controllo riveduta tenendo in considerazione i riscontri raccolti durante la fase pilota.

Il gruppo di esperti ad alto livello sull'intelligenza artificiale è indipendente ed è stato istituito dalla Commissione europea nel giugno 2018.

Referente: Nathalie Smuha - Coordinatrice del gruppo di esperti ad alto livello sull'IA
E-mail: CNECT-HLG-AI@ec.europa.eu

Commissione europea
B-1049 Bruxelles

Documento reso pubblico il 8 aprile 2019.

Un primo progetto del presente documento, pubblicato il 18 dicembre 2018, è stato sottoposto a una consultazione pubblica durante la quale oltre 500 partecipanti hanno fornito il proprio riscontro. Il gruppo desidera esprimere un vivo ringraziamento a tutti coloro che hanno fornito un riscontro sul primo progetto di documento. Tali riscontri sono stati presi in considerazione durante la preparazione della presente versione riveduta.

Ulteriori informazioni sul gruppo di esperti ad alto livello sull'intelligenza artificiale sono disponibili online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

SOMMARIO

SINTESI

(1) L'obiettivo degli orientamenti è promuovere un'IA affidabile. Un'IA affidabile si basa su **tre componenti** che dovrebbero essere presenti durante l'intero ciclo di vita del sistema: a) **legalità**, l'IA deve ottemperare a tutte le leggi e ai regolamenti applicabili, b) **eticità**, l'IA deve assicurare l'adesione a principi e valori etici, e c) **robustezza**, dal punto di vista tecnico e sociale poiché, anche con le migliori intenzioni, i sistemi di IA possono causare danni non intenzionali. Ciascuna componente in sé è necessaria ma non sufficiente per realizzare un'IA affidabile. Idealmente le tre componenti operano armonicamente e si sovrappongono; qualora, nella pratica, si dovessero verificare tensioni tra di esse la società dovrebbe adoperarsi per risolverle.

(2) I presenti orientamenti definiscono il **quadro di riferimento per realizzare un'IA affidabile**. Esso non affronta esplicitamente la prima componente (legalità dell'IA[1]), ma offre piuttosto orientamenti per promuovere e garantire l'eticità e la robustezza dell'IA (la seconda e la terza componente). I presenti orientamenti sono rivolti a tutti i portatori di interessi e intendono offrire qualcosa in più di un semplice elenco di principi etici, fornendo indicazioni su come tali principi possano essere resi operativi in sistemi sociotecnici. Essi sono forniti su tre livelli di astrazione, dal più astratto capitolo I al più concreto capitolo III, e si concludono con esempi di opportunità e di serie preoccupazioni generate dai sistemi di IA.

I. Il capitolo I, partendo da un approccio basato sui diritti fondamentali, individua i **principi etici** e i valori correlati che devono essere rispettati nello sviluppo, nella distribuzione e nell'utilizzo dei sistemi di IA.

Indicazioni chiave tratte dal capitolo I

- ✓ Sviluppare, distribuire e utilizzare sistemi di IA aderendo ai seguenti principi etici: *rispetto dell'autonomia umana, prevenzione dei danni, equità ed esplicabilità.* Riconoscere e risolvere le potenziali tensioni tra questi principi.
- ✓ Prestare particolare attenzione alle situazioni che coinvolgono gruppi più vulnerabili, come i bambini, le persone con disabilità e altri gruppi storicamente svantaggiati o a rischio di esclusione e alle situazioni caratterizzate da asimmetrie di potere o di informazione, ad esempio tra datori di lavoro e lavoratori o tra imprese e consumatori[2].
- ✓ Riconoscere e tenere presente che i sistemi di IA, pur offrendo vantaggi concreti agli individui e alla società, comportano anche dei rischi e possono avere effetti negativi, anche difficili da prevedere, individuare o misurare (ad esempio sulla democrazia, sullo Stato di diritto, sulla giustizia distributiva o sulla stessa mente umana). Se necessario, adottare provvedimenti adeguati ad attenuare tali rischi, in modo proporzionato alla loro portata.

II. Attingendo al capitolo I, il capitolo II fornisce indicazioni su come realizzare un'IA affidabile elencando **sette requisiti** che i sistemi di IA dovrebbero soddisfare e per la cui attuazione possono essere utilizzati metodi tecnici e non tecnici.

Indicazioni chiave tratte dal capitolo II

- ✓ Garantire che lo sviluppo, la distribuzione e l'utilizzo dei sistemi di IA soddisfino i requisiti di un'IA affidabile: 1) intervento e sorveglianza umani, 2) robustezza tecnica e sicurezza, 3) riservatezza e governance dei dati, 4) trasparenza, 5) diversità, non discriminazione ed equità, 6) benessere sociale e ambientale e 7) accountability.

---

[1] Tutte le affermazioni normative del presente documento intendono rispecchiare le indicazioni per il raggiungimento della seconda e terza componente di un'IA affidabile (eticità e robustezza). Pertanto, tali affermazioni non intendono fornire né pareri giuridici né indicazioni su come ottemperare alle leggi vigenti, sebbene vada riconosciuto che molte di tali affermazioni sono in qualche misura già presenti nelle leggi vigenti. A tal proposito, cfr. paragrafo 21 e seguenti.

[2] Cfr. articoli da 24 a 27 della Carta dei diritti fondamentali dell'Unione europea, che trattano dei diritti del minore, dei diritti degli anziani, dell'inserimento delle persone con disabilità e dei diritti dei lavoratori. Cfr. anche articolo 38 che tratta della protezione dei consumatori.

- ✓ Prendere in considerazione metodi tecnici e non tecnici per garantire l'attuazione di tali requisiti.
- ✓ Favorire la ricerca e l'innovazione a sostegno della valutazione dei sistemi di IA e del raggiungimento del rispetto dei requisiti; diffondere i risultati e le domande aperte al grande pubblico e formare sistematicamente una nuova generazione di esperti in etica dell'IA.
- ✓ Informare in modo chiaro e proattivo i portatori di interessi in merito alle capacità e ai limiti del sistema di IA, creando così aspettative realistiche, e in merito ai modi in cui i requisiti sono attuati. Essere trasparenti circa il fatto che si sta interagendo con un sistema di IA.
- ✓ Agevolare la tracciabilità e la verificabilità dei sistemi di IA, in particolare in contesti o situazioni critiche.
- ✓ Coinvolgere i portatori di interessi durante l'intero ciclo di vita del sistema di IA. Promuovere la formazione e l'istruzione affinché tutti i portatori di interessi siano formati e informati in merito all'IA affidabile.
- ✓ Essere consapevoli che vi potrebbero essere tensioni fondamentali tra i diversi principi e i diversi requisiti. Individuare, valutare, documentare e comunicare costantemente le soluzioni di compromesso.

III. Il capitolo III fornisce una lista di controllo concreta ma non esaustiva per la valutazione dell'affidabilità dell'IA volta a rendere operativi i requisiti enunciati nel capitolo II. Tale **lista di controllo** dovrà essere adattata agli specifici casi d'uso del sistema di IA[3].

Indicazioni chiave tratte dal capitolo III

- ✓ Adottare la lista di controllo per la valutazione dell'affidabilità dell'IA nelle fasi di sviluppo, distribuzione o utilizzo dei sistemi di IA e adattarla allo specifico caso d'uso in cui il sistema è applicato.
- ✓ Tenere presente che tale lista di controllo per la valutazione non sarà mai esaustiva. Garantire che un'IA affidabile non sia una questione di caselle da spuntare, ma un processo continuo di individuazione e attuazione dei requisiti, valutazione delle soluzioni e miglioramento dei risultati durante l'intero ciclo di vita del sistema di IA, e di coinvolgimento dei portatori di interessi in tale processo.

(3) Nella sezione finale del documento viene data forma concreta alle questioni toccate nel quadro di riferimento, offrendo esempi di opportunità vantaggiose che dovrebbero essere perseguite e di serie preoccupazioni che i sistemi di IA suscitano e che dovrebbero essere attentamente considerate.

(4) Il proposito dei presenti orientamenti è quello di offrire indicazioni relative alle applicazioni di IA in generale, realizzando così una base orizzontale per ottenere un'IA affidabile, tuttavia situazioni differenti danno vita a sfide differenti. Si dovrebbe pertanto esaminare se, oltre a tale quadro di riferimento orizzontale, non sia necessario anche un approccio settoriale, data la specificità contestuale dei sistemi di IA.

(5) I presenti orientamenti non intendono sostituire eventuali azioni politiche o di regolamentazione attuali o future, né hanno lo scopo di scoraggiarne l'adozione. Dovrebbero essere considerati come un documento vivo da rivedere e aggiornare nel corso del tempo, al fine di garantirne la costante pertinenza di pari passo con l'evoluzione della tecnologia, dei nostri ambienti sociali e della nostra conoscenza. Il presente documento è

[3] In linea con l'ambito di applicazione del quadro di riferimento descritto al paragrafo 2, la lista di controllo non fornisce orientamenti su come garantire la conformità giuridica (legalità dell'IA), ma si limita a offrire indicazioni per soddisfare la seconda e la terza componente di un'IA affidabile (eticità e robustezza dell'IA).

concepito come punto di partenza per una discussione su "un'IA affidabile per l'Europa"[4]. Al di là dell'Europa, i presenti orientamenti hanno inoltre l'obiettivo di promuovere la ricerca, la riflessione e la discussione su un quadro etico per i sistemi di IA a livello mondiale.

[4] Tale ideale è valido sia per i sistemi di IA sviluppati, distribuiti e utilizzati negli Stati membri dell'UE sia per quelli sviluppati o prodotti altrove, ma distribuiti e utilizzati nell'UE. Nel presente documento quando si parla di "Europa" si intende includere gli Stati membri dell'UE. I presenti orientamenti, tuttavia, ambiscono ad essere pertinenti anche al di fuori dell'UE. A tal proposito occorre notare che sia la Norvegia che la Svizzera fanno parte del piano coordinato sull'IA convenuto e pubblicato nel dicembre 2018 dalla Commissione e dagli Stati membri.

A. INTRODUZIONE

(6) Nelle sue comunicazioni del 25 aprile 2018 e del 7 dicembre 2018, la Commissione europea (la Commissione) ha esposto la propria visione di intelligenza artificiale (IA), a sostegno di "un'IA 'made in Europe' etica, sicura e all'avanguardia"[5]. La visione della Commissione si fonda su tre pilastri: i) aumentare gli investimenti pubblici e privati nell'IA per promuoverne l'adozione, ii) prepararsi ai cambiamenti socioeconomici e iii) garantire un quadro etico e giuridico adeguato a rafforzare i valori europei.

(7) Per sostenere l'attuazione di tale visione, la Commissione ha istituito il gruppo di esperti ad alto livello sull'intelligenza artificiale, un gruppo indipendente incaricato di elaborare due documenti: 1) gli orientamenti etici per l'IA e 2) le raccomandazioni sugli investimenti e la politica.

(8) Il presente documento contiene gli orientamenti etici per l'IA, riveduti dopo ulteriori riflessioni in seno al gruppo, alla luce dei riscontri ricevuti a seguito dalla consultazione pubblica sul progetto di documento pubblicato il 18 dicembre 2018. Il documento si basa anche sui lavori del Gruppo europeo per l'etica delle scienze e delle nuove tecnologie[6] e si ispira ad altri esercizi simili[7].

(9) Nel corso degli ultimi mesi, i 52 membri del gruppo di esperti si sono riuniti, hanno discusso e dialogato, nel rispetto del motto europeo "uniti nella diversità". Il gruppo crede che l'IA possa trasformare considerevolmente la società. L'IA non rappresenta un fine in sé, ma piuttosto un mezzo promettente per aumentare la prosperità umana, migliorando così il benessere individuale e sociale e il bene comune nonché favorendo progresso e innovazione. I sistemi di IA possono in particolare contribuire a facilitare il conseguimento degli obiettivi di sviluppo sostenibile delle Nazioni Unite, come la promozione dell'equilibrio di genere e la lotta ai cambiamenti climatici, la razionalizzazione dell'uso delle risorse naturali, il miglioramento della salute, della mobilità e dei processi produttivi e il sostegno al monitoraggio dei progressi compiuti rispetto agli indicatori di sostenibilità e coesione sociale.

(10) A tal fine si devono realizzare sistemi di IA [8]**antropocentrici**, tenendo fede all'impegno di metterli sempre al servizio dell'umanità e del bene comune, con l'obiettivo di migliorare il benessere e la libertà degli esseri umani. Pur offrendo grandi opportunità, i sistemi di IA comportano anche rischi che devono essere gestiti in modo appropriato e proporzionato. Disponiamo di un'importante occasione per plasmare il loro sviluppo, perciò intendiamo garantire che gli ambienti sociotecnici in cui sono integrati siano affidabili e che i produttori dei sistemi di IA ottengano un vantaggio competitivo integrando nei loro prodotti e servizi un'IA affidabile. A tal fine occorre **massimizzare i benefici dei sistemi di IA**, **prevenendone e minimizzandone al tempo stesso i rischi**.

(11) In un contesto di rapidi cambiamenti tecnologici, riteniamo essenziale che la fiducia rimanga il cemento delle società, delle comunità, delle economie e dello sviluppo sostenibile. Pertanto, la **nostra ambizione fondamentale è un'IA affidabile**, poiché gli esseri umani e le comunità riusciranno ad avere fiducia nello sviluppo della tecnologia e nelle sue applicazioni solo quando esisterà un quadro di riferimento chiaro e completo per conseguirne l'affidabilità.

---

5 COM(2018) 237 finale COM(2018) 795 final. Si noti che il termine "made in Europe" è utilizzato in tutta la comunicazione della Commissione. L'ambito di applicazione dei presenti orientamenti mira tuttavia a comprendere non solo i sistemi di IA prodotti in Europa, ma anche quelli sviluppati altrove e distribuiti o utilizzati in Europa. In tutto il presente documento, il proposito è quindi promuovere un'IA affidabile "per" l'Europa.

6 Il Gruppo europeo per l'etica delle scienze e delle nuove tecnologie è un gruppo consultivo della Commissione.

7 Cfr. sezione 3.3 del documento COM(2018) 237 final.

8 Una definizione di sistemi di IA ai fini del presente documento è fornita nel glossario alla fine del documento stesso. Tale definizione è approfondita in un documento elaborato appositamente dal gruppo di esperti ad alto livello sull'intelligenza artificiale e allegato ai presenti orientamenti, intitolato "Una definizione di IA: principali capacità e discipline scientifiche".

(12) A nostro avviso l'Europa dovrebbe seguire questo percorso per divenire culla e leader della tecnologia etica e all'avanguardia. Noi, come cittadini europei, attraverso un'IA affidabile potremo sfruttarne i vantaggi compatibilmente con il rispetto dei diritti umani, della democrazia e dello Stato di diritto, valori per noi fondamentali.

IA affidabile

(13) Per le persone e le società l'affidabilità rappresenta un prerequisito per lo sviluppo, la distribuzione e l'utilizzo di sistemi di IA. Se i sistemi di IA - e gli esseri umani che li creano - non sono innegabilmente degni di fiducia possono verificarsi conseguenze indesiderate e l'adozione dell'IA potrebbe essere ostacolata, impedendo la realizzazione dei benefici sociali ed economici potenzialmente enormi apportati da tali sistemi. Per aiutare l'Europa a conseguire tali benefici, crediamo sia necessario fare dell'etica un pilastro fondamentale per garantire e calibrare un'IA affidabile.

(14) La fiducia nello sviluppo, nella distribuzione e nell'utilizzo dei sistemi di IA non riguarda solo le proprietà intrinseche della tecnologia, ma anche le qualità dei sistemi sociotecnici che comportano applicazioni di IA[9]. Analogamente al problema (alla perdita) della fiducia nell'aviazione, nell'energia nucleare o nella sicurezza alimentare, non sono semplicemente le componenti del sistema di IA che possono generare o meno fiducia, ma il sistema nel suo contesto generale. L'impegno a favore di un'IA affidabile non riguarda solo l'affidabilità del sistema di IA stesso, ma richiede un approccio olistico e sistemico, che comporti l'affidabilità di tutti gli attori e di tutti i processi appartenenti al contesto sociotecnico del sistema durante l'intero ciclo di vita.

(15) Un'IA affidabile possiede **tre componenti** che dovrebbero essere sempre presenti durante l'intero ciclo di vita del sistema:

1. **legalità**, l'IA deve ottemperare a tutte le leggi e a tutti i regolamenti applicabili,
2. **eticità**, l'IA deve assicurare l'adesione a principi e valori etici, e
3. **robustezza**, dal punto di vista tecnico e sociale poiché, anche con le migliori intenzioni, i sistemi di IA possono causare danni non intenzionali.

(16) Ciascuna di queste tre componenti è necessaria ma non sufficiente in sé per realizzare un'IA affidabile[10]. Idealmente le tre componenti operano armonicamente e si sovrappongono; nella pratica, tuttavia, si possono creare tensioni tra di esse (ad esempio, a volte l'ambito di applicazione e il contenuto della legislazione vigente potrebbero non essere in linea con le norme etiche). Adoperarci affinché tutte e tre le componenti contribuiscano a garantire un'IA affidabile[11] rientra nella nostra responsabilità individuale e collettiva in quanto società.

(17) Un approccio affidabile è fondamentale per consentire una "competitività responsabile", gettando le basi affinché tutti coloro che hanno a che fare con sistemi di IA possano essere certi che la progettazione, lo sviluppo e l'utilizzo di tali sistemi sono leciti, etici e robusti. I presenti orientamenti hanno lo scopo di promuovere un'innovazione responsabile e sostenibile dell'IA in Europa. Essi mirano a fare dell'etica un pilastro fondamentale per sviluppare un approccio unico all'IA volto a favorire, rendere possibile e salvaguardare la prosperità umana a livello individuale e il bene comune a livello sociale. Riteniamo che ciò permetterà all'Europa di affermarsi come leader mondiale con un'IA all'avanguardia e degna della fiducia individuale e collettiva. Solo garantendo l'affidabilità i cittadini europei potranno trarre il massimo vantaggio dai sistemi di IA, certi del fatto che sono state adottate misure di salvaguardia contro i suoi rischi potenziali.

(18) Proprio come l'utilizzo dei sistemi di IA non si ferma alle frontiere nazionali, così anche i loro effetti; sono

---

9 Tali sistemi includono esseri umani, attori statali, imprese, infrastrutture, software, protocolli, norme, governance, diritto vigente, meccanismi di sorveglianza, strutture di incentivi, procedure di verifica, segnalazione di migliori pratiche e altro.

10 Ciò non esclude che ulteriori condizioni possano essere o diventare necessarie.

11 Ciò significa anche che il legislatore o i responsabili politici potrebbero dover rivedere l'adeguatezza della legislazione vigente qualora essa non fosse in linea con i principi etici.

quindi necessarie soluzioni globali per le opportunità e le sfide globali che accompagnano l'IA. Incoraggiamo pertanto tutti i portatori di interessi ad adoperarsi per creare un quadro di riferimento globale per un'IA affidabile, costruendo un consenso internazionale e promuovendo e sostenendo nel contempo un approccio basato sui diritti fondamentali.

Destinatari e ambito di applicazione

(19) I presenti orientamenti sono rivolti a tutti i portatori di interessi nel campo dell'IA che progettano, sviluppano, distribuiscono, implementano, utilizzano l'IA o ne sono interessati, compresi, a titolo esemplificativo, imprese, organizzazioni, ricercatori, servizi pubblici, agenzie governative, istituzioni, organizzazioni della società civile, soggetti privati, lavoratori e consumatori. I portatori di interessi impegnati a conseguire un'IA affidabile possono scegliere, se lo desiderano, di utilizzare i presenti orientamenti come metodo per rendere operativo il loro impegno, in particolare utilizzando la pratica lista di controllo del capitolo III nei loro processi di sviluppo e distribuzione dei sistemi di IA. La lista di controllo può anche completare i processi di valutazione esistenti e quindi esservi integrata.

(20) Il proposito degli orientamenti è quello di offrire indicazioni relative ad applicazioni di IA in generale, realizzando così una base orizzontale per ottenere un'IA affidabile. Tuttavia, **situazioni differenti danno vita a sfide differenti**. I sistemi di IA che offrono consigli musicali non suscitano le stesse preoccupazioni etiche dei sistemi di IA che propongono terapie mediche fondamentali. Allo stesso modo, sono differenti le opportunità e le sfide comportate dai sistemi di IA utilizzati nel contesto delle relazioni tra impresa e consumatore, tra impresa e impresa, tra datore di lavoro e lavoratore, tra settore pubblico e cittadini o, più in generale, utilizzati in diversi settori o casi d'uso. Tenuto conto della specificità contestuale dei sistemi di IA, l'attuazione dei presenti orientamenti deve essere adattata all'applicazione di IA specifica. Andrebbe inoltre esaminata la necessità di un ulteriore approccio settoriale a integrazione del quadro di riferimento orizzontale più generale proposto nel presente documento.

Per comprendere meglio come i presenti orientamenti possano essere attuati a livello orizzontale e le materie che richiedono un approccio settoriale, si invitano tutti i portatori di interessi a sperimentare la lista di controllo per la valutazione dell'affidabilità dell'IA (capitolo III) che rende operativo il presente quadro di riferimento e a fornire un riscontro. Sulla base dei riscontri raccolti durante la fase pilota, la lista di controllo contenuta nei presenti orientamenti sarà riveduta entro l'inizio del 2020. La fase pilota sarà avviata entro l'estate del 2019 e durerà fino alla fine dell'anno. Tutti i portatori di interessi che lo desiderano potranno partecipare manifestando il loro interesse tramite l'Alleanza europea per l'IA.

B. UN QUADRO DI RIFERIMENTO PER UN'IA AFFIDABILE

(21) I presenti orientamenti definiscono un quadro di riferimento per ottenere un'IA affidabile basata sui diritti fondamentali sanciti dalla Carta dei diritti fondamentali dell'Unione europea e dal pertinente diritto internazionale in materia di diritti umani. Di seguito sono illustrate brevemente le tre componenti di un'IA affidabile.

Legalità dell'IA

(22) I sistemi di IA non operano in un mondo senza leggi. A livello europeo, nazionale e internazionale un corpus normativo giuridicamente vincolante è già in vigore o è pertinente per lo sviluppo, la distribuzione e l'utilizzo dei sistemi di IA. Le fonti giuridiche pertinenti sono, a titolo esemplificativo, il diritto primario dell'UE (i trattati dell'Unione europea e la sua Carta dei diritti fondamentali), il diritto derivato dell'UE (ad esempio il regolamento generale sulla protezione dei dati, le direttive antidiscriminazione, la direttiva macchine, la direttiva sulla responsabilità dei prodotti, il regolamento sulla libera circolazione dei dati non personali, il diritto dei consumatori e le direttive in materia di salute e sicurezza sul lavoro), ma anche i trattati ONU sui

diritti umani e le convenzioni del Consiglio d'Europa (come la Convenzione europea dei diritti dell'uomo) e numerose leggi degli Stati membri dell'UE. Oltre alle norme applicabili orizzontalmente, esistono varie norme specifiche per settore applicabili a particolari applicazioni di IA (ad esempio il regolamento sui dispositivi medici nel settore sanitario).

(23) Il diritto prevede obblighi sia positivi che negativi, il che significa che l'interpretazione della legge deve avvenire non solo in riferimento a ciò che *non si può* fare, ma anche in riferimento a ciò che si *deve* fare. La legge proibisce certe azioni e ne consente altre. A tal proposito, è opportuno osservare che la Carta dell'UE contiene articoli sulla "libertà d'impresa" e sulla "libertà delle arti e delle scienze", oltre ad articoli riguardanti settori più conosciuti nell'ambito della garanzia dell'affidabilità dell'IA, come ad esempio la protezione dei dati e la non discriminazione.

(24) Gli orientamenti non trattano esplicitamente la prima componente dell'IA affidabile (legalità dell'IA), ma offrono piuttosto indicazioni per promuovere e garantire la seconda e terza componente (eticità e robustezza dell'IA). Queste ultime sono già presenti in una certa misura nel diritto vigente, ma la loro piena realizzazione può andare al di là degli obblighi giuridici esistenti.

(25) Nulla di quanto contenuto nel presente documento deve essere inteso o interpretato come consulenza legale o indicazioni su come conformarsi alle norme e alle disposizioni giuridiche vigenti. Nulla di quanto contenuto nel presente documento crea diritti giuridici o impone obblighi giuridici nei confronti di terzi. Si rammenta tuttavia che è dovere di ogni persona fisica o giuridica rispettare le leggi, sia che esse siano attualmente applicabili sia che siano adottate in futuro in base allo sviluppo dell'IA. I presenti orientamenti partono dal presupposto che **tutti i diritti e gli obblighi giuridici che si applicano ai processi e alle attività di sviluppo, distribuzione e utilizzo dell'IA restano inderogabili e devono essere debitamente rispettati.**

Eticità dell'IA

(26) Per ottenere un'IA affidabile non è sufficiente il rispetto della legge, che è solo una delle tre componenti. Il diritto non è sempre al passo con gli sviluppi tecnologici, e a volte non lo è nemmeno con le norme etiche o semplicemente non è adatto ad affrontare determinate questioni. Affinché i sistemi di IA siano affidabili, essi dovrebbero quindi essere anche etici garantendo la compatibilità con le norme etiche.

Robustezza dell'IA

(27) Anche qualora il fine etico sia garantito, gli individui e la società devono comunque essere sicuri che i sistemi di IA non causeranno alcun danno involontario. Tali sistemi dovrebbero funzionare in modo sicuro e affidabile e dovrebbero essere previste misure di salvaguardia per prevenire qualsiasi effetto negativo indesiderato. È quindi importante garantire che i sistemi di IA siano robusti. Tale componente è necessaria sia da un punto di vista tecnico (garantendo la robustezza tecnica del sistema in un dato contesto, ad esempio il settore di applicazione o la fase del ciclo di vita), sia da un punto di vista sociale (tenendo in debita considerazione il contesto e l'ambiente in cui il sistema opera). L'eticità e la robustezza dell'IA sono quindi componenti strettamente correlate che si integrano a vicenda. I principi enunciati nel capitolo I e i requisiti tratti da essi nel capitolo II riguardano entrambe le componenti.

Il quadro di riferimento

(28) Nel presente documento le indicazioni sono fornite su tre livelli di astrazione, dal più astratto capitolo I al più concreto capitolo III:

I) Basi di un'IA affidabile. Il capitolo I getta le basi di un'IA affidabile, delineando l'approccio basato sui diritti fondamentali[12]. Identifica e descrive i principi etici che devono essere rispettati al fine di garantire l'eticità e la

[12] I diritti fondamentali sono alla base del diritto internazionale e dell'Unione in materia di diritti umani e dei diritti giuridicamente tutelati garantiti dai trattati UE e dalla Carta dei diritti fondamentali dell'UE. Essendo giuridicamente vincolante, il rispetto dei diritti fondamentali rientra quindi nella prima componente di un'IA affidabile: la legalità dell'IA. I diritti fondamentali possono tuttavia essere intesi

robustezza dell'IA.

II) Realizzazione di un'IA affidabile. Il capitolo II traduce questi principi etici in sette requisiti che i sistemi di IA dovrebbero attuare e soddisfare durante l'intero ciclo della loro vita e offre metodi tecnici e non tecnici che possono essere utilizzati per la loro attuazione.

III) Valutazione di un'IA affidabile. Gli operatori del settore dell'IA si attendono orientamenti concreti. Il capitolo III contiene pertanto una lista di controllo per la valutazione dell'affidabilità dell'IA, preliminare e non esaustiva, per rendere operativi i requisiti del capitolo II. Tale valutazione va adattata alla specifica applicazione del sistema.

(29) Nella sezione finale del documento sono illustrate le opportunità vantaggiose dell'IA e le serie preoccupazioni che essa suscita, che non vanno trascurate e sulle quali desideriamo incoraggiare un'ulteriore discussione.

(30) La struttura degli orientamenti è illustrata nella *figura 1*.

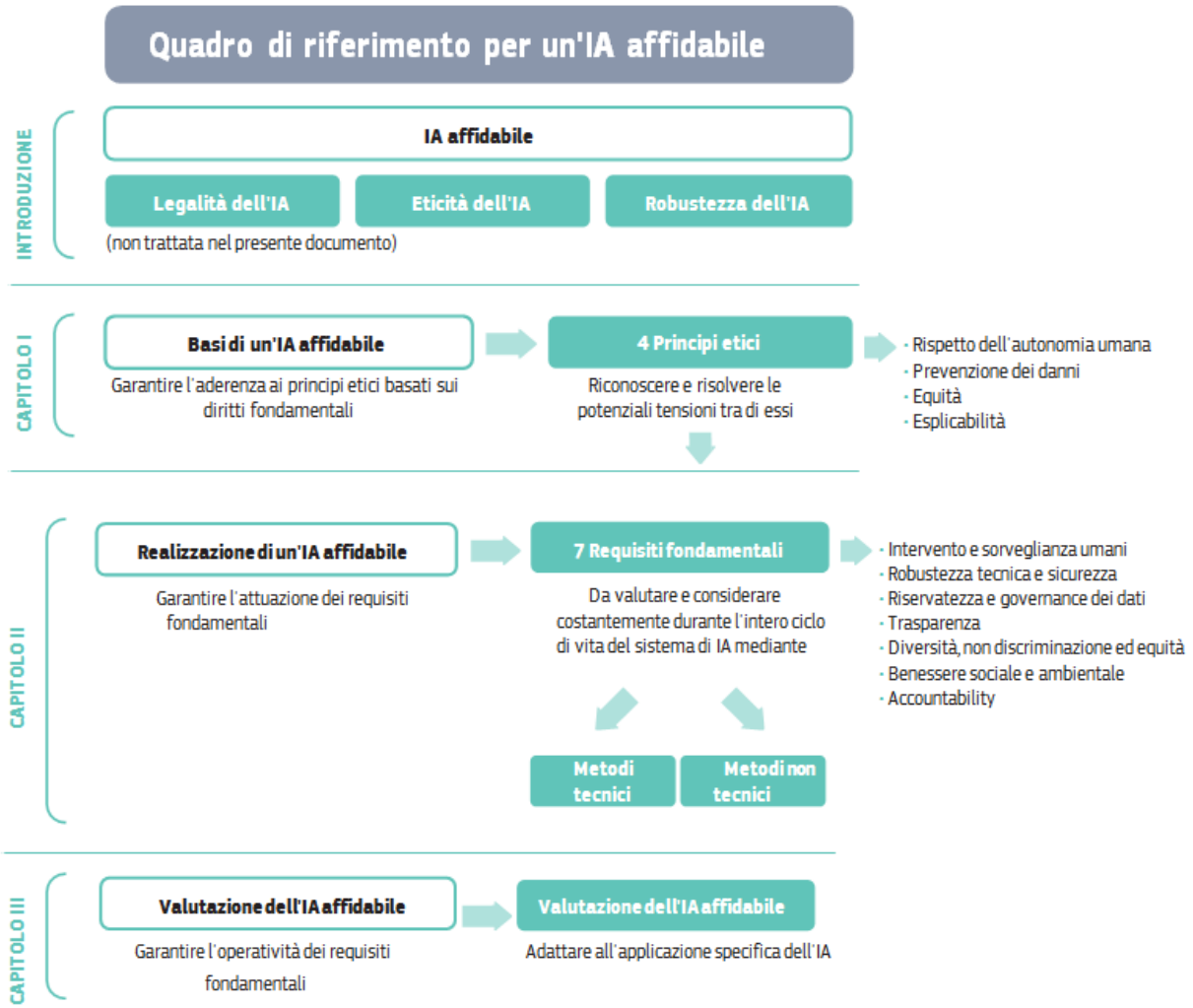

Figura 1: gli orientamenti come quadro di riferimento per un'IA affidabile

anche come il riflesso di diritti morali speciali che tutti gli individui posseggono in virtù della loro umanità, indipendentemente dal loro status giuridicamente vincolante. In tal senso, fanno parte anche della seconda componente dell'IA affidabile: l'eticità dell'IA.

I. Capitolo I - Basi di un'IA affidabile

(31) Nel presente capitolo si delineano le basi di un'IA affidabile fondata sui diritti fondamentali e riflesse nei quattro principi etici che dovrebbero essere rispettati per garantire l'eticità e la robustezza dell'IA. Il capitolo attinge ampiamente alla sfera dell'etica.

(32) L'etica dell'IA è una branca dell'etica applicata che studia gli interrogativi etici posti dallo sviluppo, dalla distribuzione e dall'utilizzo dell'IA. Il suo interesse principale risiede nell'individuare come l'IA possa favorire o mettere a rischio la felicità degli individui, sia in termini di qualità della vita che di autonomia umana e libertà necessarie per una società democratica.

(33) La riflessione etica sulla tecnologia dell'IA può essere utile per vari scopi. In primo luogo, può stimolare la riflessione sulla necessità di proteggere individui e gruppi al livello più elementare. In secondo luogo, può stimolare nuovi tipi di innovazione che promuovano valori etici, come quelli che contribuiscono al raggiungimento degli obiettivi di sviluppo sostenibile delle Nazioni Unite[13], che sono parte integrante della Agenda 2030 dell'UE di imminente pubblicazione[14]. Sebbene il presente documento riguardi principalmente il primo scopo citato, non va sottovalutata l'importanza che l'etica potrebbe avere per il secondo. Un'IA affidabile può migliorare la prosperità individuale e il benessere collettivo generando agiatezza, creando valore e massimizzando la ricchezza. Può contribuire alla realizzazione di una società equa, in quanto può aiutare ad accrescere la salute e il benessere dei cittadini secondo modalità che promuovono l'uguaglianza nella distribuzione delle opportunità economiche, sociali e politiche.

(34) È quindi indispensabile capire come sostenere nel miglior modo possibile lo sviluppo, la distribuzione e l'utilizzo dell'IA, al fine di garantire che tutti possano prosperare in un mondo basato sull'IA e di costruire un futuro migliore rimanendo competitivi a livello globale. L'uso nella nostra società dei sistemi di IA, come di qualsiasi tecnologia potente, pone vari problemi etici, ad esempio in merito al loro effetto sulle persone e sulla società, sulle capacità decisionali e sull'incolumità. Se ricorreremo sempre di più all'assistenza dei sistemi di IA o se delegheremo loro le decisioni dobbiamo essere certi, tramite adeguati processi di accountability, che gli effetti di tali sistemi sulla vita delle persone siano equi, che rispettino valori intransigibili e che siano in grado di agire di conseguenza.

(35) L'Europa deve definire quale visione normativa intende realizzare per un futuro immerso nell'IA e di conseguenza deve capire quale nozione di IA dovrebbe essere studiata, sviluppata, distribuita e utilizzata in Europa per concretizzare tale visione. Con il presente documento, intendiamo contribuire a tale sforzo introducendo la nozione di IA affidabile, che riteniamo sia il modo giusto per costruire un futuro con l'IA. Un futuro in cui la democrazia, lo Stato di diritto e i diritti fondamentali sono alla base dei sistemi di IA e in cui tali sistemi migliorano e difendono costantemente la cultura democratica consentirà anche di creare un ambiente in cui l'innovazione e la competitività responsabile possono prosperare.

(36) Un codice etico specifico per settore, per quanto sia coerente, sviluppato e perfezionato in successive versioni, non potrà mai sostituire il ragionamento etico stesso, il quale deve sempre rimanere sensibile ai dettagli contestuali che non possono essere racchiusi in orientamenti generali. Oltre a sviluppare un corpus normativo, per garantire un'IA affidabile occorre costruire e mantenere una cultura e una mentalità etiche attraverso il dibattito pubblico, l'istruzione e l'apprendimento pratico.

1. I diritti fondamentali come titolarità di diritti morali e giuridici

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_it.

14 https://sustainabledevelopment.un.org/?menu=1300.

(37) Crediamo in un approccio all'etica dell'IA basato sui diritti fondamentali sanciti dai trattati UE,[15] dalla Carta dei diritti fondamentali dell'Unione europea e dal diritto internazionale in materia di diritti umani[16]. Il rispetto dei diritti fondamentali, nel quadro della democrazia e dello Stato di diritto, costituisce la base più incoraggiante per individuare principi e valori etici astratti che possono essere resi operativi nei sistemi di IA.

(38) I trattati dell'UE e la Carta dell'UE sanciscono una serie di diritti fondamentali che gli Stati membri e le istituzioni dell'UE sono giuridicamente tenuti a rispettare quando attuano il diritto unionale. Tali diritti sono descritti nella Carta dei diritti fondamentali dell'Unione europea con riferimento alla dignità, alle libertà, all'uguaglianza, alla solidarietà, ai diritti dei cittadini e alla giustizia. Il fondamento che accomuna questi diritti può essere inteso come radicato nel rispetto della dignità umana, riflettendo così quello che definiamo un "approccio antropocentrico" in cui l'essere umano gode di uno status morale unico e inalienabile di primato in campo civile, politico, economico e sociale.[17]

(39) Sebbene i diritti sanciti dalla Carta dei diritti fondamentali dell'Unione europea siano giuridicamente vincolanti[18], è importante riconoscere che i diritti fondamentali non sempre contemplano una tutela giuridica completa. Per la Carta dei diritti fondamentali dell'Unione europea, ad esempio, è importante sottolineare che il suo ambito di applicazione è limitato alle materie del diritto unionale. Il diritto internazionale in materia di diritti umani e in particolare la Convenzione europea dei diritti dell'uomo sono giuridicamente vincolanti per gli Stati membri dell'UE, anche in materie che esulano dall'ambito di applicazione del diritto unionale. Al contempo va sottolineato che i diritti fondamentali sono conferiti anche agli individui e (in certa misura) ai gruppi in virtù del loro status morale di esseri umani, indipendentemente dalla loro forza giuridica. Intesi come diritti giuridicamente applicabili, i diritti fondamentali rientrano pertanto nella prima componente di un'IA affidabile (legalità dell'IA), che salvaguarda il rispetto della legge. Intesi come diritti di ciascuno, radicati nello status morale intrinseco degli esseri umani, essi sono anche alla base della seconda componente dell'IA affidabile (eticità dell'IA), che si occupa di norme etiche che non sono necessariamente vincolanti dal punto di vista giuridico ma cruciali per garantire l'affidabilità. Poiché il presente documento non intende offrire indicazioni sulla prima componente, ai fini dei presenti orientamenti non vincolanti, i riferimenti ai diritti fondamentali riflettono la seconda componente.

2. Dai diritti fondamentali ai principi etici

2.1 I diritti fondamentali come base per un'IA affidabile

(40) Nell'esauriente corpus di diritti indivisibili previsti dal diritto internazionale in materia di diritti umani, dai trattati UE e dalla Carta dei diritti fondamentali dell'Unione europea, le seguenti famiglie di diritti fondamentali sono particolarmente pertinenti per quanto riguarda i sistemi di IA. Molti di questi diritti, in determinate circostanze, sono giuridicamente applicabili nell'UE, pertanto il rispetto dei loro termini è giuridicamente vincolante. Ma anche una volta rispettati i diritti fondamentali giuridicamente applicabili, la riflessione etica può aiutarci a comprendere come lo sviluppo, la distribuzione e l'utilizzo dell'IA possano implicare i diritti fondamentali e i valori soggiacenti, e può contribuire a fornire un orientamento più dettagliato nel percorso di ricerca di ciò che *dobbiamo* fare piuttosto che ciò che (attualmente) *possiamo* fare con la tecnologia.

---

[15] L'UE si fonda su un impegno costituzionale a tutelare i diritti fondamentali e indivisibili degli esseri umani, a garantire il rispetto dello Stato di diritto, a promuovere la libertà democratica e il bene comune. Questi diritti sono sanciti negli articoli 2 e 3 del trattato sull'Unione europea e nella Carta dei diritti fondamentali dell'UE.

[16] Tali impegni sono ulteriormente specificati in altri strumenti giuridici, quali ad esempio la Carta sociale europea del Consiglio d'Europa o in atti legislativi specifici come il regolamento generale dell'UE sulla protezione dei dati.

[17] Occorre notare che l'impegno per un'IA antropocentrica e il suo nesso con i diritti fondamentali necessita di basi collettive e costituzionali in cui la libertà individuale e il rispetto della dignità umana siano praticabili e significative, anziché implicare un indebito tornaconto individualistico dell'essere umano.

[18] A norma dell'articolo 51, le disposizioni della Carta si applicano alle istituzioni dell'Unione come pure agli Stati membri nell'attuazione del diritto dell'Unione.

(41) ***Rispetto della dignità umana***. Il concetto di dignità umana racchiude l'idea che ogni essere umano possiede un "valore intrinseco", che non deve mai essere svilito, compromesso o represso dagli altri e nemmeno dalle nuove tecnologie come i sistemi di IA[19]. Nel contesto dell'IA, il rispetto per la dignità umana implica che tutte le persone siano trattate con il rispetto loro dovuto in quanto *soggetti* morali, piuttosto che come semplici *oggetti* da vagliare, catalogare, valutare per punteggio, aggregare, condizionare o manipolare. I sistemi di IA devono quindi essere sviluppati in modo che rispettino, servano e proteggano l'integrità fisica e psichica degli esseri umani, il senso di identità personale e culturale e la soddisfazione dei bisogni essenziali.[20]

(42) ***Libertà individuale***. Gli esseri umani devono rimanere liberi di prendere decisioni importanti per se stessi. Ciò comporta la libertà dall'intrusione di organismi sovrani, ma richiede anche l'intervento di organizzazioni governative e non governative per garantire che individui o popolazioni a rischio di esclusione abbiano pari accesso ai benefici e alle opportunità offerti dall'IA. Nell'ambito dell'IA, per salvaguardare la libertà individuale occorre ridurre al minimo la coercizione illegittima diretta o indiretta, le minacce all'autonomia mentale e alla salute psichica, la sorveglianza ingiustificata, l'inganno e la manipolazione iniqua. La libertà individuale comporta di fatto un impegno affinché gli individui possano esercitare un controllo addirittura maggiore sulla propria vita, il che include (tra gli altri diritti) la tutela della libertà d'impresa, della libertà delle arti e delle scienze, della libertà di espressione, del diritto alla vita privata e alla riservatezza, della libertà di riunione e di associazione**.**

(43) ***Rispetto della democrazia, della giustizia e dello Stato di diritto***. Tutti i poteri dello Stato nelle democrazie costituzionali devono essere giuridicamente autorizzati e limitati dalla legge. I sistemi di IA devono servire a mantenere e a promuovere i processi democratici e a rispettare la pluralità dei valori e delle scelte di vita degli individui. Essi non devono compromettere i processi democratici, la decisione umana o i sistemi di voto democratico. Nei sistemi di IA deve essere insito l'impegno a garantire di non operare con modalità che compromettano gli impegni di base su cui si fonda lo Stato di diritto, le leggi e i regolamenti obbligatori e a garantire il giusto processo e l'uguaglianza di fronte alla legge.

(44) ***Uguaglianza, non discriminazione e solidarietà*** (*compresi i diritti delle persone a rischio di esclusione*). Si deve garantire pari rispetto per il valore morale e la dignità di tutti gli esseri umani. Ciò va oltre la non discriminazione, che tollera la distinzione tra situazioni diverse sulla base di giustificazioni oggettive. In un contesto di IA, l'uguaglianza implica che il funzionamento del sistema non possa generare risultati ingiustamente distorti (ad esempio, i dati utilizzati per istruire i sistemi di IA dovrebbero essere il più inclusivi possibile e rappresentare gruppi di popolazione diversi). Ciò richiede anche l'adeguato rispetto per le persone e i gruppi potenzialmente vulnerabili[21], come i lavoratori, le donne, le persone con disabilità, le minoranze etniche, i bambini, i consumatori o altri soggetti a rischio di esclusione.

(45) ***Diritti dei cittadini***. I cittadini godono di un'ampia gamma di diritti, tra cui il diritto di voto, il diritto a una buona amministrazione o all'accesso ai documenti pubblici e il diritto di presentare petizioni all'amministrazione. I sistemi di IA possono sostanzialmente migliorare la portata e l'efficienza della fornitura di beni e servizi pubblici alla società da parte dei governi ma, allo stesso tempo, le applicazioni di IA potrebbero avere effetti negativi sui diritti dei cittadini che dovrebbero essere salvaguardati. Utilizzando il termine "diritti dei cittadini" non si negano né trascurano i diritti dei cittadini di paesi terzi e delle persone irregolari (o illegali) presenti nell'UE che sono tra l'altro tutelati dal diritto internazionale e quindi godono di diritti anche nel campo dell'IA.

---

19 C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights, *EJIL*, 19(4), 2008.

20 Per una comprensione della "dignità umana" in questo senso si veda E. Hilgendorf, Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity, in: D. Grimm, A. Kemmerer, C. Möllers (a cura di), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, pag. 325 ss.

21 Per una definizione del termine, nel senso utilizzato nel presente documento, si veda il glossario.

2.2 Principi etici nel contesto dei sistemi di IA[22]

(46) Molte organizzazioni pubbliche, private e civili hanno tratto ispirazione dai diritti fondamentali per produrre quadri etici per l'IA[23]. Nell'UE, il Gruppo europeo per l'etica delle scienze e delle nuove tecnologie ha proposto una serie di 9 principi di base, basati sui valori fondamentali sanciti dai trattati e dalla Carta dei diritti fondamentali dell'Unione europea[24]. Consolidiamo tale lavoro riconoscendo la maggior parte dei principi finora proposti dai vari gruppi e chiarendo al contempo le finalità che tutti i principi cercano di alimentare e sostenere. Tali principi etici possono ispirare nuovi e specifici strumenti normativi, possono contribuire a interpretare i diritti fondamentali in funzione dell'evoluzione del nostro ambiente sociotecnico e possono orientare la logica alla base dello sviluppo, dell'utilizzo e dell'implementazione dei sistemi di IA, adattandosi dinamicamente all'evoluzione della società stessa.

(47) I sistemi di IA devono migliorare il benessere individuale e collettivo. Questa sezione elenca **quattro principi etici**, radicati nei diritti fondamentali ai quali occorre aderire per garantire che i sistemi di IA siano sviluppati, distribuiti e utilizzati in modo affidabile. Sono definiti come **imperativi etici** affinché gli operatori del settore dell'IA si adoperino sempre per aderirvi. Senza imporre alcuna gerarchia, i principi sono elencati di seguito secondo l'ordine di apparizione dei diritti fondamentali su cui si basano all'interno della Carta dei diritti fondamentali dell'Unione europea.[25]

(48) I principi sono i seguenti:

i) rispetto dell'autonomia umana
ii) prevenzione dei danni
iii) equità
iv) esplicabilità

(49) Molti di questi sono in larga misura già presenti in disposizioni giuridiche vigenti che devono essere obbligatoriamente ottemperate e quindi rientrano anche nell'ambito di applicazione della prima componente dell'IA affidabile, ovvero la legalità[26]. Tuttavia, come indicato sopra, anche se molti obblighi giuridici riflettono principi etici, l'adesione ai principi etici va oltre il rispetto formale del diritto vigente.[27]

- Il principio del rispetto dell'autonomia umana

(50) I diritti fondamentali su cui si fonda l'Unione europea sono volti a garantire il rispetto della libertà e dell'autonomia degli esseri umani. Gli esseri umani che interagiscono con i sistemi di IA devono poter mantenere la propria piena ed effettiva autodeterminazione e devono poter essere partecipi del processo democratico. I sistemi di IA non devono subordinare, costringere, ingannare, manipolare, condizionare o aggregare in modo ingiustificato gli esseri umani. Al contrario, devono essere progettati per aumentare, integrare e potenziare le abilità cognitive, sociali e culturali umane. La distribuzione delle funzioni tra esseri umani e sistemi di IA dovrebbe seguire i principi di progettazione antropocentrica e lasciare ampie opportunità

[22] Tali principi si applicano anche allo sviluppo, alla distribuzione e all'utilizzo di altre tecnologie, e quindi non sono peculiari dei sistemi di IA. Il presente paragrafo tenta di esporne la pertinenza specifica in un contesto relativo all'IA.

[23] Il ricorso ai diritti fondamentali contribuisce inoltre a limitare l'incertezza normativa, in quanto può basarsi su decenni di pratica di tutela dei diritti fondamentali nell'UE, offrendo in tal modo chiarezza, leggibilità e prevedibilità.

[24] Recentemente, la task force di AI4People ha esaminato i summenzionati principi del Gruppo europeo per l'etica delle scienze e delle nuove tecnologie più altri 36 principi etici presentati finora e li ha classificati in quattro grandi categorie: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), "*AI4People —An Ethical Framework for a Good IA Society: Opportunities, Risks, Principles, and Recommendations*", Minds and Machines 28(4): 689-707.

[25] Il rispetto dell'autonomia umana è strettamente connesso al diritto alla dignità umana e alla libertà (sanciti dagli articoli 1 e 6 della Carta). La prevenzione dei danni è strettamente connessa alla protezione dell'integrità fisica e psichica (sancita dall'articolo 3). L'equità è strettamente connessa ai diritti alla non discriminazione, alla solidarietà e alla giustizia (sanciti dall'articolo 21 e seguenti). L'esplicabilità e la responsabilità sono strettamente connesse ai diritti relativi alla giustizia (sanciti dall'articolo 47).

[26] Si pensi ad esempio al regolamento generale sulla protezione dei dati o alle norme UE per la tutela dei consumatori.

[27] Per approfondimenti dell'argomento, si veda ad esempio L. Floridi, Soft Ethics and the Governance of the Digital, *Philosophy & Technology,* marzo 2018, volume 31, numero 1, pagg. 1-8.

di scelta all'essere umano. Ciò significa garantire la sorveglianza[28] e il controllo dei processi operativi nei sistemi di IA da parte di esseri umani. I sistemi di IA possono anche cambiare radicalmente il mondo del lavoro, sostenendo l'uomo nell'ambiente lavorativo al fine di generare un lavoro significativo.

- Il principio della prevenzione dei danni

(51) I sistemi di IA non devono causare danni[29] né aggravarli e neppure influenzare negativamente gli esseri umani[30], per cui occorre tutelare la dignità umana nonché l'integrità fisica e psichica. I sistemi di IA e gli ambienti in cui operano devono essere sicuri e protetti. Devono essere tecnicamente robusti e si deve garantire che non siano esposti ad usi malevoli. Le persone vulnerabili dovrebbero ricevere maggiore attenzione ed essere incluse nello sviluppo e nella distribuzione dei sistemi di IA. Occorre prestare particolare attenzione anche alle situazioni in cui i sistemi di IA possono causare o aggravare gli effetti negativi dovuti ad asimmetrie di potere o di informazione, come ad esempio tra datori di lavoro e dipendenti, imprese e consumatori o governi e cittadini. La prevenzione dei danni implica anche il rispetto dell'ambiente naturale e di tutti gli esseri viventi.

- Il principio di equità

(52) Lo sviluppo, la distribuzione e l'utilizzo dei sistemi di IA devono essere equi. Pur riconoscendo che esistono molte interpretazioni di equità, noi reputiamo che l'equità abbia una dimensione sia sostanziale che procedurale. La dimensione sostanziale implica un impegno a garantire una distribuzione giusta ed equa di costi e di benefici e a garantire che gli individui e i gruppi siano liberi da distorsioni inique, discriminazioni e stigmatizzazioni. Riuscendo a evitare distorsioni inique, i sistemi di IA potrebbero persino aumentare l'equità sociale. Occorre inoltre promuovere le pari opportunità in termini di accesso all'istruzione, ai beni, ai servizi e alla tecnologia. L'utilizzo dei sistemi di IA, inoltre, non deve mai ingannare gli utenti (finali) né ostacolarne la libertà di scelta. Inoltre, l'equità implica che gli operatori del settore dell'IA rispettino il principio di proporzionalità tra mezzi e fini, e valutino attentamente come bilanciare interessi e obiettivi concorrenti.[31] La dimensione procedurale dell'equità implica la capacità di impugnare le decisioni elaborate dai sistemi di IA e dagli esseri umani che li gestiscono[32]e la possibilità di presentare un ricorso efficace contro di esse. A tal fine, l'organismo responsabile della decisione deve essere identificabile e i processi decisionali devono essere spiegabili.

- Il principio dell'esplicabilità

(53) L'esplicabilità è fondamentale per creare e mantenere la fiducia degli utenti nei sistemi di IA. Tale principio implica che i processi devono essere trasparenti, le capacità e lo scopo dei sistemi di IA devono essere comunicati apertamente e le decisioni, per quanto possibile, devono poter essere spiegate a coloro che ne sono direttamente o indirettamente interessati. Senza tali informazioni, una decisione non può essere debitamente impugnata. Non sempre è possibile spiegare, tuttavia, perché un modello ha generato un particolare risultato o decisione (e quale combinazione di fattori di input vi ha contribuito). È il cosiddetto caso della "scatola nera" i cui algoritmi richiedono un'attenzione particolare. In tali circostanze, possono essere

---

28 I l concetto di sorveglianza umana è approfondito al seguente paragrafo 65.

29 I danni possono essere individuali o collettivi e possono includere danni immateriali all'ambiente sociale, culturale e politico.

30 Ciò comprende anche il modo di vivere degli individui e dei gruppi sociali, evitando ad esempio danni culturali.

31 Ciò è correlato al principio di proporzionalità (racchiuso nella massima secondo cui non si deve "sparare ai passeri con i cannoni"). Le misure adottate per raggiungere un fine (ad esempio, le misure di estrazione dei dati attuate per realizzare la funzione di ottimizzazione dell'IA) dovrebbero essere limitate allo stretto necessario. Ciò implica anche che, quando più misure competono per la soddisfazione di un fine, si dovrebbe dare la preferenza a quella meno lesiva dei diritti fondamentali e delle norme etiche (ad esempio, gli sviluppatori di IA dovrebbero sempre preferire i dati del settore pubblico ai dati personali). È possibile inoltre far riferimento alla proporzionalità tra utente e distributore, considerando i diritti delle imprese (compresa la proprietà intellettuale e la riservatezza), da un lato, e i diritti dell'utente, dall'altro.

32 Anche facendo ricorso al loro diritto di associazione e di adesione a un sindacato in un ambiente di lavoro, come previsto dall'articolo 12 della Carta dei diritti fondamentali dell'Unione europea.

necessarie altre misure per garantire l'esplicabilità (ad esempio, la tracciabilità, la verificabilità e la comunicazione trasparente sulle capacità del sistema), posto che il sistema nel suo complesso rispetti i diritti fondamentali. Il grado di esplicabilità necessario dipende in larga misura dal contesto e dalla gravità delle conseguenze nel caso in cui il risultato sia errato o comunque impreciso.[33]

2.3 Tensioni tra i principi

(54) Tra i principi summenzionati possono insorgere tensioni e non esiste alcuna soluzione prestabilita per risolverle. Per affrontare tali tensioni si dovrebbero definire metodi di discussione responsabile, coerenti con l'impegno fondamentale dell'UE a favore della partecipazione democratica, del processo equo e di una partecipazione politica aperta. Ad esempio, in vari settori di applicazione, il *principio di prevenzione dei danni* e il *principio dell'autonomia umana* possono entrare in conflitto. Si consideri l'esempio dell'uso di sistemi di IA nell'ambito della "polizia predittiva" che può contribuire a ridurre la criminalità, ma con modalità di sorveglianza che interferiscono con la libertà individuale e la riservatezza. I benefici complessivi dei sistemi di IA inoltre dovrebbero superare considerevolmente i rischi individuali prevedibili. Questi principi, pur indicando senza dubbio la direzione verso possibili soluzioni, rimangono prescrizioni etiche astratte. Non si può quindi presumere che gli operatori del settore dell'IA trovino la soluzione giusta sulla base dei principi enunciati sopra; tuttavia dovrebbero affrontare i dilemmi e i compromessi etici riflettendo razionalmente in base alle prove piuttosto che affidarsi all'intuizione o a decisioni casuali. Può accadere tuttavia in certe situazioni che non sia possibile individuare compromessi eticamente accettabili. Alcuni diritti fondamentali e i principi ad essi correlati sono assoluti e intransigibili (ad esempio, la dignità umana).

Indicazioni chiave tratte dal capitolo I

- ✓ Sviluppare, distribuire e utilizzare sistemi di IA aderendo ai seguenti principi etici: *rispetto dell'autonomia umana, prevenzione dei danni, equità ed esplicabilità.* Riconoscere e risolvere le potenziali tensioni tra questi principi.
- ✓ Prestare particolare attenzione alle situazioni che coinvolgono gruppi più vulnerabili, come i bambini, le persone con disabilità e altri gruppi storicamente svantaggiati o a rischio di esclusione e alle situazioni caratterizzate da asimmetrie di potere o di informazione, ad esempio tra datori di lavoro e lavoratori, o tra imprese e consumatori[34].
- ✓ Riconoscere e tenere presente che, pur avendo le potenzialità per offrire molti vantaggi concreti agli individui e alla società, alcune applicazioni di IA possono anche avere effetti negativi, ivi inclusi effetti difficili da prevedere, individuare o misurare (ad esempio sulla democrazia, sullo Stato di diritto, sulla giustizia distributiva o sulla stessa mente umana.) Se necessario, adottare provvedimenti adeguati ad attenuare tali rischi, in modo proporzionato alla loro portata.

II. Capitolo II - Realizzare un'IA affidabile

(55) Il presente capitolo fornisce indicazioni sull'implementazione e sulla realizzazione di un'IA affidabile basata sui principi delineati nel capitolo I, avvalendosi di un elenco di sette requisiti che dovrebbero essere soddisfatti. Vengono inoltre presentati i metodi tecnici e non tecnici attualmente disponibili per l'attuazione di questi requisiti durante l'intero ciclo di vita del sistema di IA.

1. Requisiti di un'IA affidabile

[33] Ad esempio, se le raccomandazioni di acquisto generate da un sistema di IA sono imprecise ciò non suscita grandi preoccupazioni etiche, mentre diversa è la situazione quando i sistemi di IA devono valutare l'opportunità o meno di accordare a una persona condannata a pena detentiva la libertà condizionale.

[34] Cfr. articoli da 24 a 27 della Carta dei diritti fondamentali dell'Unione europea, che trattano dei diritti del minore, dei diritti degli anziani, dell'inserimento delle persone con disabilità e dei diritti dei lavoratori. Cfr. anche articolo 38 che tratta della protezione dei consumatori.

(56) Per ottenere un'IA affidabile, i principi delineati nel capitolo I devono essere tradotti in requisiti concreti. Tali requisiti sono applicabili ai diversi portatori di interessi che partecipano al ciclo di vita dei sistemi di IA: sviluppatori, distributori, utenti finali e società in generale. Per sviluppatori si intendono i ricercatori e coloro che progettano e/o sviluppano sistemi di IA. Per distributori si intendono le organizzazioni pubbliche o private che utilizzano sistemi di IA sia all'interno dei loro processi aziendali sia per offrire prodotti e servizi a terzi. Gli utenti finali sono i soggetti coinvolti direttamente o indirettamente in un sistema di IA Infine, la società in generale comprende tutte le altre persone direttamente o indirettamente interessate dai sistemi di IA.

(57) Le diverse categorie di portatori di interessi rivestono diversi ruoli nel garantire che i requisiti siano soddisfatti:

a. gli sviluppatori attuano e applicano i requisiti ai processi di progettazione e sviluppo;
b. i distributori garantiscono che i sistemi che utilizzano e i prodotti e i servizi che offrono soddisfino i requisiti;
c. gli utenti finali e la società in generale sono informati su questi requisiti e hanno la facoltà di domandarne il rispetto.

(58) Il seguente elenco di requisiti non è esaustivo.[35] Comprende aspetti sistemici, individuali e sociali:

1 Intervento e sorveglianza umani
Inclusi i diritti fondamentali, l'intervento umano e la sorveglianza umana

2 Robustezza tecnica e sicurezza
Inclusi la resilienza agli attacchi e la sicurezza, il piano di emergenza e la sicurezza generale, la precisione, l'affidabilità e la riproducibilità.

3 Riservatezza e governance dei dati
Inclusi il rispetto della riservatezza, la qualità e l'integrità dei dati e l'accesso ai dati.

4 Trasparenza
Incluse la tracciabilità, la spiegabilità e la comunicazione

5 Diversità, non discriminazione ed equità
Incluse la prevenzione di distorsioni inique, l'accessibilità e la progettazione universale, e la partecipazione dei portatori di interessi

6 Benessere sociale e ambientale
Inclusi la sostenibilità e il rispetto ambientale, l'impatto sociale, la società e la democrazia

7 Accountability
Inclusi la verificabilità, la riduzione al minimo degli effetti negativi e la loro segnalazione, i compromessi e i ricorsi.

[35] Senza imporre alcuna gerarchia, i principi sono elencati di seguito secondo l'ordine di apparizione dei principi e dei diritti cui fanno riferimento all'interno della Carta dei diritti fondamentali dell'Unione europea.

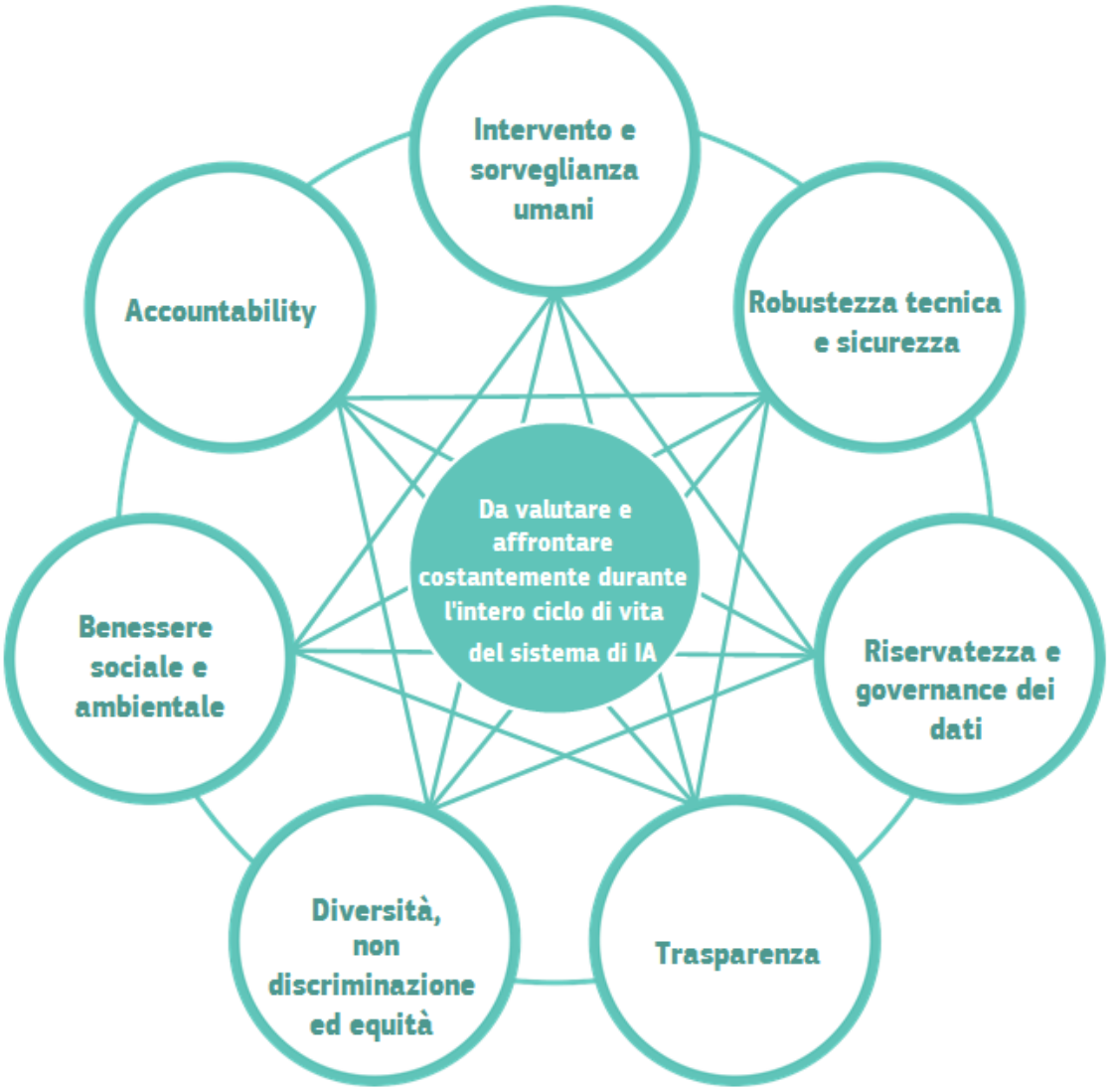

Figura 2: interrelazione dei sette requisiti. Sono tutti di pari importanza, si avvalorano vicendevolmente e dovrebbero essere attuati e valutati durante l'intero ciclo di vita di un sistema di IA.

(59) Sebbene i requisiti siano tutti di pari importanza, al momento di applicarli in diversi campi e settori occorrerà tenere in considerazione il contesto e le potenziali tensioni che possono insorgere tra essi. L'attuazione di tali requisiti, che dovrebbe avvenire durante l'intero ciclo di vita di un sistema di IA, dipende dall'applicazione specifica. La maggior parte dei requisiti si applica a tutti i sistemi di IA, tuttavia è dedicata particolare attenzione a quelli che influiscono direttamente o indirettamente sugli individui. Per alcune applicazioni (ad esempio in ambienti industriali), essi possono avere pertanto minore pertinenza.

(60) I requisiti summenzionati comprendono elementi che in alcuni casi compaiono già nel diritto vigente. Ribadiamo che, in linea con la prima componente di un'IA affidabile, spetta agli sviluppatori e ai distributori dei sistemi di IA garantire il rispetto degli obblighi giuridici, sia per quanto riguarda le norme ad applicazione orizzontale sia per quanto riguarda la regolamentazione specifica di settore.

(61) Ciascun requisito verrà approfondito nei paragrafi seguenti.

1. <u>Intervento e sorveglianza umani</u>

(62) I sistemi di IA dovrebbero sostenere l'autonomia e il processo decisionale umani, come prescrive il principio del *rispetto dell'autonomia umana*. A tal fine essi devono agire come catalizzatori di una società democratica, prospera ed equa, sostenendo l'intervento degli utenti e promuovendo i diritti fondamentali, e devono consentire la sorveglianza umana.

(63) ***Diritti fondamentali***. Come molte tecnologie, anche i sistemi di IA possono in egual misura agevolare e ostacolare i diritti fondamentali. Possono rappresentare un vantaggio per le persone: ad esempio quando le aiutano a tener traccia dei loro dati personali o quando consentono loro un maggior accesso all'istruzione, sostenendo così il diritto all'istruzione. Tuttavia, data la loro portata e capacità, i sistemi di IA possono anche influire negativamente sui diritti fondamentali. In situazioni in cui esistono rischi di questo tipo, dovrebbe essere effettuata una valutazione d'impatto sui diritti fondamentali prima dello sviluppo del sistema, includendo una valutazione della possibilità di ridurre i rischi o di giustificarli in quanto necessari in una società democratica, al fine di rispettare i diritti e le libertà altrui. Dovrebbero essere, inoltre, messi in atto meccanismi per ottenere riscontri esterni sui sistemi di IA che potenzialmente violano i diritti fondamentali.

(64) ***Intervento umano.*** Gli utenti dovrebbero essere in grado di adottare decisioni autonome e informate in merito ai sistemi di IA. Dovrebbero ricevere le conoscenze e gli strumenti per comprendere e interagire con i sistemi di IA a un livello soddisfacente e, ove possibile, essere in grado di valutare autonomamente o contestare il sistema in modo ragionevole. I sistemi di IA dovrebbero aiutare gli individui a compiere scelte migliori e più informate, coerenti con i loro obiettivi. I sistemi di IA talvolta possono essere utilizzati per plasmare e influenzare il comportamento umano attraverso meccanismi difficili da individuare, in quanto possono sfruttare processi subconsci, quali ad esempio varie forme di manipolazione iniqua, l'inganno, la frode, l'aggregazione e il condizionamento, tutti meccanismi che possono minacciare l'autonomia individuale. Il principio generale dell'autonomia dell'utente deve essere al centro della funzionalità del sistema. A tal fine è fondamentale il diritto di non essere sottoposti a una decisione basata unicamente sul trattamento automatizzato quando questa produca effetti giuridici sugli utenti o qualora incida in modo analogo significativamente su di loro.[36]

(65) ***Sorveglianza umana***. La sorveglianza umana aiuta a garantire che un sistema di IA non comprometta l'autonomia umana o provochi altri effetti negativi. La sorveglianza può avvenire mediante meccanismi di governance che consentano un approccio con intervento umano (human-in-the-loop - HITL), con supervisione umana (human-on-the-loop - HOTL) o con controllo umano (human-in-command - HIC). L'approccio HITL prevede la possibilità di intervento umano in ogni ciclo decisionale del sistema, che in molti casi non è né possibile né auspicabile. L'approccio HOTL prevede l'intervento umano durante il ciclo di progettazione del sistema e il monitoraggio del funzionamento del sistema. L'approccio HIC prevede il controllo dell'attività del sistema di IA nel suo complesso (compresi i suoi effetti generali a livello economico, sociale, giuridico ed etico) e la capacità di decidere quando e come utilizzare il sistema in qualsiasi particolare situazione. Si potrebbe anche decidere di non utilizzare un sistema di IA in una data situazione, di stabilire livelli di discrezionalità umana durante l'uso del sistema, o di garantire la capacità di ignorare una decisione presa da un sistema. Occorre inoltre garantire che le autorità pubbliche competenti abbiano la capacità di esercitare la sorveglianza in conformità al loro mandato. Potrebbero essere necessari meccanismi di sorveglianza a vari livelli a sostegno di altre misure di sicurezza e di controllo, a seconda del settore di applicazione del sistema di IA e del rischio potenziale. A parità di condizioni, minore è la sorveglianza che un essere umano può esercitare su un sistema di IA, maggiore è la necessità di prove esaurienti e di una governance rigorosa.

2. Robustezza tecnica e sicurezza

(66) Una componente cruciale per ottenere un'IA affidabile è la robustezza tecnica, che è strettamente connessa al

[36] È possibile far riferimento all'articolo 22 del regolamento generale sulla protezione dei dati in cui è già sancito tale diritto.

principio di prevenzione dei danni. Per garantire la robustezza tecnica è necessario che i sistemi di IA siano sviluppati con un approccio di prevenzione dei rischi e in maniera tale che si comportino in maniera attendibile secondo le previsioni, riducendo al minimo i danni non intenzionali e imprevisti e prevenendo danni inaccettabili. Ciò dovrebbe valere anche per i potenziali cambiamenti nel loro ambiente operativo o per la presenza di altri agenti (umani e artificiali) che possono interagire con il sistema in maniera contraddittoria. Si dovrebbe inoltre garantire l'integrità fisica e psichica degli esseri umani.

(67) ***Resilienza agli attacchi e sicurezza***. I sistemi di IA, come tutti i sistemi software, dovrebbero essere protetti contro le vulnerabilità che li espongono allo sfruttamento da parte degli avversari, ad esempio l'hacking. Gli attacchi possono colpire i dati (avvelenamento dei dati), il modello (*model leakage*) o l'infrastruttura sottostante, sia software che hardware. Se il sistema di IA è attaccato, ad esempio con input ingannevoli (*adversarial attack*), possono essere modificati sia i dati che il comportamento del sistema, il quale può essere indotto ad adottare decisioni diverse o ad arrestarsi completamente. Il danneggiamento di sistemi e dati può anche essere frutto di intenzioni malevole o dell'esposizione a situazioni impreviste e processi di sicurezza inadeguati possono anche sfociare in decisioni errate o perfino in danni fisici. Per poter considerare sicuri i sistemi di IA,[37] si dovrebbe tener conto delle possibili applicazioni non intenzionali di IA (ad esempio applicazioni a duplice uso) e di potenziali abusi di un sistema di IA da parte di soggetti malintenzionati e si dovrebbero adottare misure per prevenirli e mitigarli.[38]

(68) ***Piano di emergenza e sicurezza generale***. I sistemi di IA dovrebbero essere dotati di misure di salvaguardia che attivino un piano di emergenza in caso di problemi. I sistemi di IA potrebbero quindi passare da una procedura statistica a una procedura basata su regole, oppure richiedere un operatore umano prima di continuare la loro azione.[39] Occorre garantire che il sistema faccia ciò che è tenuto a fare senza danneggiare gli esseri viventi o l'ambiente, compresa la riduzione al minimo delle conseguenze e degli errori non intenzionali. Si dovrebbero inoltre mettere in atto processi per chiarire e valutare i rischi potenziali associati all'uso dei sistemi di IA in vari campi di applicazione. Il livello delle misure di sicurezza dipende dall'entità del rischio posto da un sistema di IA, che a sua volta dipende dalle capacità del sistema. Nei casi in cui è possibile prevedere che il processo di sviluppo o il sistema stesso comporteranno rischi particolarmente elevati, è fondamentale sviluppare e testare le misure di sicurezza in modo proattivo.

(69) ***Precisione.*** La precisione attiene alla capacità di un sistema di IA di produrre un giudizio corretto, ad esempio per classificare correttamente le informazioni nelle categorie appropriate, o la sua capacità di fare previsioni, formulare raccomandazioni o adottare decisioni esatte sulla base di dati o modelli. Un processo di sviluppo e valutazione esplicito e ben strutturato può sostenere, mitigare e correggere i rischi non intenzionali derivanti da previsioni imprecise e, quando non è possibile evitare tali imprecisioni occasionali, è importante che il sistema sia in grado di indicare la probabilità di errore. Un alto livello di precisione è particolarmente importante in situazioni in cui il sistema di IA influisce direttamente sulle vite umane.

(70) ***Affidabilità e riproducibilità.*** È fondamentale che i risultati dei sistemi di IA siano riproducibili e affidabili. Un sistema IA è affidabile se funziona correttamente con una serie di input e in varie situazioni. Ciò è necessario per esaminare un sistema di IA e prevenire danni involontari. La riproducibilità indica se un esperimento di IA mostra lo stesso comportamento quando ripetuto nelle stesse condizioni. Ciò consente agli scienziati e ai responsabili politici di descrivere accuratamente quello che fanno i sistemi di IA. I file di replica[40] possono

---

37 Cfr. ad esempio le considerazioni al punto 2.7 del piano coordinato sull'intelligenza artificiale dell'Unione europea.

38 Per la sicurezza dei sistemi di IA, potrebbe essere decisamente necessario sviluppare un circolo virtuoso in ricerca e sviluppo tra la comprensione degli attacchi, lo sviluppo di adeguate protezioni e il miglioramento delle metodologie di valutazione. Per raggiungere questo obiettivo, dovrebbe essere promossa una convergenza tra la comunità dell'IA e la comunità della sicurezza. Inoltre, è responsabilità di tutti i soggetti interessati elaborare norme comuni di sicurezza transfrontaliere e creare un ambiente di fiducia reciproca, favorendo la collaborazione internazionale. Per le misure possibili, cfr. "*Malicious Use of IA*" (Avin S., Brundage M., et. al., 2018).

39 Si dovrebbero considerare anche degli scenari in cui l'intervento umano non sia immediatamente possibile.

40 Si tratta di file che replicano ogni fase del processo di sviluppo del sistema di IA, dalla ricerca e dalla raccolta dei dati iniziali fino ai risultati.

facilitare il processo di prova e riproduzione dei comportamenti.

3. Riservatezza e governance dei dati

(71) La riservatezza, strettamente connessa al *principio di prevenzione dei danni*, è un diritto fondamentale particolarmente interessato dall'IA. Per prevenire danni alla riservatezza occorre tra l'altro un'adeguata governance dei dati che riguardi la qualità e l'integrità dei dati utilizzati, la loro pertinenza rispetto al settore in cui i sistemi di IA saranno distribuiti, i protocolli di accesso e la capacità di trattare i dati in modo da tutelare la riservatezza.

(72) ***Riservatezza e protezione dei dati***. I sistemi di IA devono garantire la riservatezza e la protezione dei dati durante l'intero ciclo di vita del sistema[41], comprese le informazioni fornite inizialmente dall'utente e quelle che lo riguardano, nonché le informazioni sull'utente generate nel corso della sua interazione con il sistema (ad esempio, gli output generati dal sistema di IA per utenti specifici o le modalità di risposta degli utenti a particolari raccomandazioni). Le registrazioni digitali del comportamento umano possono permettere ai sistemi di IA di dedurre non solo le preferenze individuali, ma anche il loro orientamento sessuale, l'età, il genere, le opinioni religiose o politiche. Affinché le persone abbiano fiducia nel processo di raccolta dei dati che le riguardano, occorre garantire che tali dati non siano utilizzati ai fini di un'illecita o iniqua discriminazione.

(73) ***Qualità e integrità dei dati***. La qualità dei set di dati utilizzati è di fondamentale importanza per le prestazioni dei sistemi di IA. I dati che vengono raccolti, possono contenere distorsioni, imprecisioni, errori e sbagli socialmente costruiti, ed è un aspetto da affrontare prima di addestrare la macchina con un determinato set di dati. Si deve inoltre garantire l'integrità dei dati. Se si immettono dati malevoli, un sistema di IA può cambiare il suo comportamento, in particolare con i sistemi di autoapprendimento. I processi e i set di dati utilizzati devono essere testati e documentati in ogni fase (ad es. pianificazione, addestramento, prova e distribuzione). Ciò dovrebbe valere anche per i sistemi di IA che non sono sviluppati internamente ma acquisiti altrove.

(74) ***Accesso ai dati***. In ogni organizzazione che gestisce i dati personali (che si tratti o meno di un utente del sistema), dovrebbero essere messi in atto protocolli di dati che ne regolino l'accesso. Tali protocolli dovrebbero indicare chi può accedere ai dati e in quali circostanze. Solo il personale debitamente qualificato con la competenza e la necessità di accedere ai dati personali dovrebbe essere autorizzato a farlo.

4. Trasparenza

(75) Questo requisito è strettamente connesso al *principio dell'esplicabilità* e comprende la trasparenza degli elementi pertinenti per un sistema di IA: i dati, il sistema e i modelli di business.

(76) ***Tracciabilità***. I set di dati e i processi che determinano la decisione del sistema di IA, compresi quelli di raccolta ed etichettatura dei dati, come pure gli algoritmi utilizzati, dovrebbero essere documentati secondo i migliori standard per consentire la tracciabilità e aumentare la trasparenza. Ciò vale anche per le decisioni prese dal sistema di IA, in quanto tale documentazione consente di capire perché un sistema di IA ha preso una decisione errata e, di conseguenza, potrebbe aiutare a prevenire errori futuri. La tracciabilità facilita quindi la verificabilità e la spiegabilità.

(77) ***Spiegabilità.*** La spiegabilità attiene alla capacità di spiegare sia i processi tecnici di un sistema di IA che le relative decisioni umane (ad esempio i settori di applicazione di un sistema di IA). Affinché un sistema di IA possa essere tecnicamente spiegabile gli esseri umani devono poter capire e tenere traccia delle decisioni prese dal sistema stesso. Potrebbe inoltre essere necessario trovare un compromesso tra il miglioramento

[41] È possibile citare le leggi vigenti in materia di riservatezza, come il regolamento generale sulla protezione dei dati o il prossimo regolamento sulla e-privacy.

della spiegabilità di un sistema (sacrificando la precisione) e l'aumento della precisione (a scapito della spiegabilità). Se un sistema di IA influisce considerevolmente sulla vita delle persone, dovrebbe sempre essere possibile richiedere una spiegazione adeguata del processo decisionale del sistema. Tale spiegazione dovrebbe essere tempestiva e adeguata alle competenze del portatore di interesse in questione (un non esperto, un'autorità di regolamentazione o un ricercatore). Dovrebbero inoltre essere disponibili indicazioni sul grado in cui un sistema di IA influenza e plasma il processo decisionale organizzativo, sulle scelte progettuali del sistema e sulla logica alla base della sua distribuzione (garantendo così la trasparenza del modello di business).

(78) ***Comunicazione.*** I sistemi di IA non devono presentarsi agli utenti come esseri umani e gli esseri umani hanno il diritto di essere a conoscenza del fatto che stanno interagendo con un sistema di IA. Ciò implica che i sistemi di IA debbano essere identificabili come tali. Inoltre, per garantire il rispetto dei diritti fondamentali, dovrebbe essere prevista ove necessario la possibilità di preferire l'interazione umana a quella con il sistema di IA. Oltre a ciò, dovrebbero essere comunicate agli operatori del settore dell'IA o agli utenti finali le capacità e le limitazioni del sistema in maniera consona al caso d'uso in questione. Ciò potrebbe comprendere la comunicazione del livello di precisione del sistema di IA e dei suoi limiti.

5. Diversità, non discriminazione ed equità

(79) Per ottenere un'IA affidabile, occorre che l'inclusione e la diversità siano permesse durante l'intero ciclo di vita del sistema. Oltre al fatto che tutti i portatori di interessi influenzati dall'IA devono essere presi in considerazione e coinvolti nel corso del processo, tale principio comporta anche la necessità di garantire la parità di trattamento e la parità di accesso attraverso processi di progettazione inclusivi. Questo requisito è strettamente connesso al *principio di equità*.

(80) ***Evitare distorsioni inique***. I set di dati utilizzati dai sistemi di IA (sia per l'addestramento che per il funzionamento) possono subire l'influenza di distorsioni storiche non intenzionali, dell'incompletezza e di modelli di cattiva governance. Se tali distorsioni permangono, determinati gruppi o persone potrebbero essere involontariamente oggetto di pregiudizi e discriminazioni[42] dirette e indirette e ciò potrebbe aggravare il pregiudizio e l'emarginazione. Il danno può anche derivare dallo sfruttamento intenzionale di distorsioni (del consumatore) o dal praticare una concorrenza sleale, con mezzi quali l'omogeneizzazione dei prezzi tramite la collusione o un mercato non trasparente.[43]Le distorsioni identificabili e discriminatorie dovrebbero essere eliminate, se possibile, nella fase di raccolta. Anche il modo in cui vengono sviluppati i sistemi di IA (ad esempio, la programmazione degli algoritmi) può subire l'influenza delle distorsioni inique. La soluzione a tal riguardo potrebbe risiedere nell'attuazione di processi di sorveglianza per analizzare e affrontare in modo chiaro e trasparente le finalità, i vincoli, i requisiti e le decisioni del sistema. L'assunzione di personale proveniente da contesti, culture e discipline diverse inoltre può garantire la diversità di opinioni e dovrebbe essere incoraggiata.

(81) ***Accessibilità e progettazione universale***. I sistemi dovrebbero essere incentrati sull'utente, soprattutto in settori relativi ai rapporti impresa-consumatore, e progettati in modo che tutte le persone possano utilizzare prodotti o servizi di IA, indipendentemente dall'età, dal genere, dalle abilità o dalle caratteristiche personali. Particolare importanza riveste l'accessibilità a questa tecnologia per le persone con disabilità, che sono presenti in tutti i gruppi sociali. I sistemi di IA non dovrebbero essere caratterizzati da un approccio a "taglia

[42] Per una definizione di discriminazione diretta e indiretta, cfr. ad esempio l'articolo 2 della direttiva 2000/78/CE del Consiglio, del 27 novembre 2000, che stabilisce un quadro generale per la parità di trattamento in materia di occupazione e di condizioni di lavoro. Cfr. anche l'articolo 21 della Carta dei diritti fondamentali dell'UE.

[43] Cfr. il documento dell'Agenzia dell'Unione europea per i diritti fondamentali: "*BigData: Discrimination in data-supported decision making (2018)*", http://fra.europa.eu/en/publication/2018/big-data-discrimination.

unica" e dovrebbero dare il dovuto peso a principi di progettazione universale[44] rivolti al più ampio spettro possibile di utenti, seguendo le norme di accessibilità pertinenti[45]. Ciò consentirà l'accesso equo e la partecipazione attiva di tutte le persone alle nuove ed emergenti attività umane mediate da computer, e con riferimento alle tecnologie assistive.[46]

(82) ***Partecipazione dei portatori di interessi***. Al fine di sviluppare sistemi di IA affidabili, è consigliabile consultare i portatori di interessi che possono essere direttamente o indirettamente interessati dal sistema durante l'intero ciclo di vita di quest'ultimo. È utile chiedere un riscontro regolare anche dopo la distribuzione e mettere in atto meccanismi a più lungo termine per la partecipazione dei portatori di interessi, ad esempio garantendo l'informazione, la consultazione e la partecipazione dei lavoratori durante l'intero processo di implementazione dei sistemi di IA presso le organizzazioni.

6. Benessere sociale e ambientale

(83) In linea con i *principi di equità* e di *prevenzione dei danni*, anche la società in generale, altri esseri senzienti e l'ambiente dovrebbero essere considerati come portatori di interessi durante l'intero ciclo di vita del sistema di IA. La sostenibilità e la responsabilità ecologica dei sistemi di IA dovrebbe essere incoraggiata e si dovrebbe promuovere la ricerca di soluzioni di IA che affrontino settori di interesse globale, come ad esempio gli obiettivi di sviluppo sostenibile. Idealmente, l'IA dovrebbe essere utilizzata a vantaggio di tutti gli esseri umani, comprese le generazioni future.

(84) ***IA sostenibile e rispettosa dell'ambiente***. Sebbene il contributo dei sistemi di IA alla soluzione di alcuni dei problemi sociali più preoccupanti sia promettente, si deve garantire che ciò avvenga il più possibile nel rispetto dell'ambiente. Il processo di sviluppo, distribuzione e utilizzo del sistema, così come l'intera catena di approvvigionamento, dovrebbero essere valutati secondo questa prospettiva, ad esempio tramite un esame critico dell'uso delle risorse e del consumo energetico durante la fase di addestramento, scegliendo le opzioni meno dannose e incoraggiando le misure che garantiscano la compatibilità ambientale dell'intera catena di approvvigionamento del sistema di IA.

(85) ***Impatto sociale***. L'onnipresente esposizione ai sistemi di IA sociale[47] in tutti gli ambiti della nostra vita (che si tratti di istruzione, lavoro, assistenza o intrattenimento) può alterare la nostra concezione di intervento sociale, o influenzare le nostre relazioni sociali e i nostri legami affettivi. I sistemi di IA possono essere utilizzati per migliorare le abilità sociali[48], ma possono parimenti contribuire al loro deterioramento, anche influendo sul benessere fisico e psichico delle persone. Gli effetti di questi sistemi devono pertanto essere attentamente monitorati e valutati.

(86) ***Società e democrazia.*** Oltre a valutare gli effetti dello sviluppo, della distribuzione e dell'utilizzo di un sistema di IA sugli individui, occorrerebbe valutare anche l'impatto sociale, tenendo conto degli effetti del sistema sulle istituzioni, sulla democrazia e sulla società in generale. L'uso dei sistemi di IA dovrebbe essere valutato attentamente, in particolare in situazioni riguardanti il processo democratico, includendo non solo quello decisionale politico, ma anche i contesti elettorali.

---

44 L'articolo 42 della direttiva sugli appalti pubblici impone che le specifiche tecniche tengano conto dei criteri di accessibilità e progettazione adeguata a tutti gli utenti.

45 Ad esempio EN 301 549.

46 Questo requisito è connesso alla Convenzione delle Nazioni Unite sui diritti delle persone con disabilità.

47 Si intendono sistemi di IA che comunicano e interagiscono con gli esseri umani simulando la socialità nell'interazione di robot umanoidi (IA incarnata) o come avatar nella realtà virtuale. In tal modo, questi sistemi possono potenzialmente modificare le nostre pratiche socioculturali e il tessuto della nostra vita sociale.

48 Cfr. ad esempio il progetto finanziato dall'UE per lo sviluppo di un software basato sull'IA che consente ai robot di interagire più efficacemente con bambini autistici in sedute di terapia condotte da esseri umani e che contribuisce a migliorare le loro capacità sociali e di comunicazione: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

7. Accountability

(87) Questo requisito integra quelli enunciati sopra ed è strettamente connesso con il *principio di equità*. Per conseguire tale requisito occorre mettere in atto meccanismi che garantiscano l'accountability dei sistemi di IA e dei loro risultati, sia prima che dopo la loro attuazione.

(88) ***Verificabilità.*** La verificabilità comporta la possibilità di valutare algoritmi, dati e processi di progettazione. Ciò non implica necessariamente che le informazioni sui modelli di business e sulla proprietà intellettuale relative al sistema di IA debbano essere sempre disponibili in modo aperto. La valutazione da parte di revisori interni ed esterni e la disponibilità delle relazioni di tale valutazione può contribuire all'affidabilità della tecnologia. Nelle applicazioni che influiscono sui diritti fondamentali, comprese le applicazioni essenziali ai fini della sicurezza, i sistemi di IA dovrebbero poter essere sottoposti a una verifica indipendente.

(89) ***Riduzione al minimo degli effetti negativi e relativa segnalazione***. Occorre garantire sia la possibilità di riferire in merito ad azioni o decisioni che contribuiscono a un determinato risultato del sistema, sia la possibilità di rispondere alle conseguenze di tale risultato. È particolarmente importante per tutti coloro che sono direttamente o indirettamente interessati identificare, valutare, riferire e ridurre al minimo i potenziali effetti negativi dei sistemi di IA. Occorre garantire la debita protezione a chi segnala irregolarità, alle ONG, ai sindacati o ad altri organismi che riferiscono in merito a legittimi timori relativi a un sistema basato sull'IA. L'uso di valutazioni d'impatto (ad esempio, red teaming o forme di valutazione d'impatto algoritmica) sia prima che durante lo sviluppo, la distribuzione e l'utilizzo di sistemi di IA può essere utile per ridurre al minimo l'impatto negativo. Tali valutazioni devono essere proporzionate al rischio posto dai sistemi di IA.

(90) ***Compromessi*** Nell'attuazione dei requisiti summenzionati possono insorgere tensioni tra i requisiti stessi, il che può condurre a inevitabili compromessi che dovrebbero essere affrontati in modo razionale e metodologico nell'ambito dello stato dell'arte. Ciò implica l'identificazione degli interessi e dei valori pertinenti coinvolti nel sistema di IA e, in caso di conflitto, occorre riconoscere esplicitamente e valutare i compromessi tra di essi in termini di rischio per i principi etici, diritti fondamentali compresi. Nelle situazioni in cui non è possibile trovare compromessi eticamente accettabili, lo sviluppo, la distribuzione e l'utilizzo del sistema di IA non dovrebbero procedere secondo quel modello. Qualunque decisione sul compromesso da accettare deve essere motivata e adeguatamente documentata. Il decisore deve assumersi la responsabilità delle modalità di attuazione del compromesso più adeguato e deve riesaminare costantemente l'adeguatezza della decisione che ne deriva per garantire che possano essere apportate le necessarie modifiche al sistema ove opportuno. [49]

(91) ***Ricorso***.Si dovrebbero prevedere meccanismi accessibili e adeguati di ricorso in caso di effetti negativi ingiusti.[50] La certezza della possibilità di ricorso in caso di esiti avversi è fondamentale per garantire la fiducia. Si dovrebbe prestare particolare attenzione alle persone o ai gruppi vulnerabili.

2. Metodi tecnici e non tecnici per realizzare un'IA affidabile

(92) Per attuare i requisiti summenzionati è possibile utilizzare metodi tecnici e non tecnici in tutte le fasi del ciclo di vita di un sistema di IA. La valutazione di tali metodi, come pure la rendicontazione e la giustificazione [51]delle modifiche ai processi di attuazione, dovrebbero avvenire su base continuativa. Poiché i sistemi di IA

[49] Per raggiungere questo obiettivo è possibile avvalersi di diversi modelli di governance. Ad esempio, la presenza di un esperto o di un comitato etico (specifico per settore) interno e/o esterno potrebbe essere utile per evidenziare settori di potenziale conflitto e suggerire le migliori soluzioni. È utile anche la consultazione e la discussione con i portatori di interessi, compresi coloro che sono esposti al rischio di essere influenzati negativamente dal sistema di IA. Le università europee dovrebbero assumere un ruolo guida nella formazione degli esperti di etica necessari.

[50] Cfr. anche il parere dell'Agenzia dell'Unione europea per i diritti fondamentali "*Improving access to remedy in the area of business and human rights at the EU level*" (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Ad esempio la giustificazione delle scelte operate nella progettazione, nello sviluppo e nella distribuzione del sistema al fine di incorporare i requisiti summenzionati.

sono in costante evoluzione e agiscono in un ambiente dinamico, la realizzazione di un'IA affidabile è un processo continuo, illustrato nella figura 3.

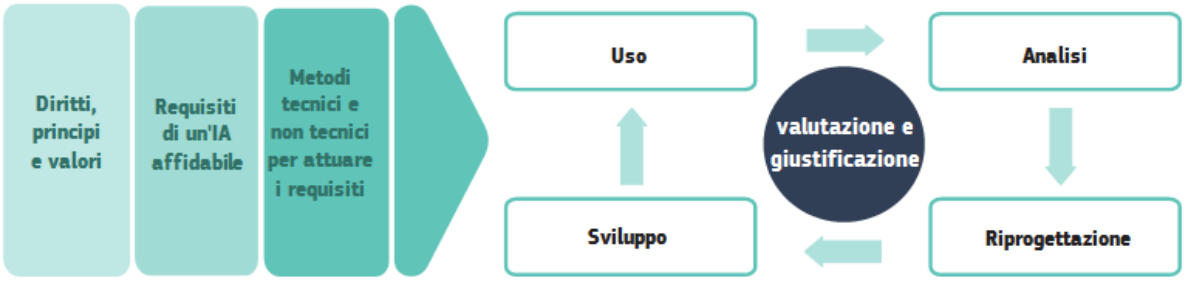

Figura 3: realizzazione di un'IA affidabile durante l'intero ciclo di vita del sistema

(93) I seguenti metodi si possono considerare reciprocamente complementari o alternativi, poiché requisiti diversi (e sensibilità diverse) possono richiedere metodi di attuazione diversi. La presente panoramica non intende essere né assoluta né esaustiva e tanto meno obbligatoria, essa mira piuttosto a offrire un elenco di proposte metodologiche che possono contribuire a implementare un'IA affidabile.

1. Metodi tecnici

(94) La presente sezione descrive i metodi tecnici per garantire un'IA affidabile che possono essere incorporati nelle fasi di progettazione, sviluppo e utilizzo di un sistema di IA. I metodi di seguito elencati variano per livello di maturità.[52]

- *Architetture per un'IA affidabile*

(95) I requisiti per un'IA affidabile dovrebbero essere "tradotti" in procedure e/o vincoli di procedura fissati nell'architettura del sistema di IA. A tal fine si potrebbe definire una "lista bianca" di regole (comportamenti o stati) che il sistema deve sempre seguire, una "lista nera" di restrizioni sui comportamenti o stati cui il sistema non deve mai trasgredire e combinazioni di queste oppure di garanzie dimostrabili più complesse che riguardano il comportamento del sistema. Il monitoraggio del rispetto di tali restrizioni durante il funzionamento del sistema può avvenire tramite un processo separato.

(96) I sistemi di IA con capacità di apprendimento che sono in grado di adattare dinamicamente il loro comportamento possono essere intesi come sistemi non deterministici che potrebbero produrre un comportamento imprevisto. Spesso tali sistemi sono valutati secondo la prospettiva teorica del ciclo "percezione/pianificazione/azione". Per adattare questa architettura affinché garantisca un'IA affidabile è necessario integrare i requisiti in tutte e tre le fasi del ciclo: i) nella fase di "percezione" si dovrebbe sviluppare il sistema in modo tale che riconosca tutti gli elementi ambientali indispensabili a garantire l'aderenza ai requisiti; ii) nella fase di "pianificazione" il sistema dovrebbe considerare solo i piani che aderiscono ai requisiti; iii) nella fase di "azione", le azioni del sistema dovrebbero essere limitate ai comportamenti che realizzano i requisiti.

(97) L'architettura descritta sopra è generica e fornisce solo una descrizione imperfetta della maggior parte dei sistemi di IA, tuttavia definisce punti di riferimento fissi per restrizioni e strategie che dovrebbero riflettersi in

[52] Mentre alcuni di questi metodi sono già disponibili attualmente, altri richiedono ancora ulteriori ricerche. I settori in cui sono necessarie ulteriori ricerche caratterizzeranno anche il secondo documento del gruppo di esperti di alto livello sull'intelligenza artificiale, vale a dire le raccomandazioni politiche e di investimento.

moduli specifici da cui deriva un sistema complessivo affidabile e percepito come tale.

- *Etica e Stato di diritto fin dalla progettazione (X-by-design)*

(98) I metodi per garantire valori fin dalla progettazione (*by-design*) forniscono legami precisi ed espliciti tra i principi astratti ai quali il sistema deve aderire e le specifiche decisioni di attuazione. L'idea che la conformità alle norme può essere implementata nella progettazione del sistema di IA rappresenta il fulcro di questo metodo. Spetta alle imprese la responsabilità di individuare fin dall'inizio sia gli effetti dei loro sistemi di IA che le norme alle quali tali sistemi dovrebbero attenersi per evitare effetti negativi. Sono già ampiamente utilizzati diversi concetti "by-design", ad esempio *privacy-by-design* (riservatezza fin dalla progettazione) e *security-by-design* (sicurezza fin dalla progettazione). Come indicato sopra, affinché l'IA ottenga fiducia, i processi, i dati e i risultati devono essere sicuri e la sua progettazione deve essere tale da resistere a dati e attacchi ingannevoli. Il sistema di IA dovrebbe implementare un meccanismo di arresto *fail safe* che consenta il riavvio dopo un arresto forzato (ad esempio un attacco).

- *Metodi esplicativi*

(99) Affinché un sistema sia affidabile, occorre essere in grado di capire perché si è comportato in un certo modo e perché ha fornito una data interpretazione. Un intero campo di ricerca, Explainable IA (XAI) cerca di affrontare questo problema per comprendere meglio i meccanismi alla base del sistema e trovare soluzioni. Attualmente, questa è ancora una sfida aperta per i sistemi di IA basati su reti neurali. Da processi di addestramento con reti neurali possono risultare parametri di rete impostati su valori numerici difficilmente correlabili con i risultati. Inoltre, a volte, piccoli cambiamenti nei valori dei dati possono concludersi in cambiamenti radicali di interpretazione, portando il sistema a confondere, ad esempio, uno scuolabus con uno struzzo. Questa vulnerabilità può essere sfruttata anche negli attacchi al sistema. I metodi XAI sono essenziali non solo per spiegare il comportamento del sistema agli utenti, ma anche per distribuire una tecnologia affidabile.

- *Prova e convalida*

(100) Data la natura non deterministica e contestualmente specifica dei sistemi di IA, i metodi di prova tradizionali non sono sufficienti. Gli errori di concetto e di rappresentazione utilizzati dal sistema possono manifestarsi solo quando un programma è applicato a dati sufficientemente realistici. Di conseguenza, per verificare e convalidare l'elaborazione dei dati, la stabilità, la robustezza e il funzionamento del modello soggiacente devono essere attentamente monitorati sia durante l'addestramento sia durante la distribuzione, entro limiti ben compresi e prevedibili. Occorre garantire che l'esito del processo di pianificazione sia coerente con gli input e che le decisioni siano adottate in modo da consentire la convalida del processo soggiacente.

(101) L'attività di prova e convalida del sistema dovrebbe essere svolta il prima possibile per assicurare che il sistema si comporti come previsto per l'intero ciclo di vita e soprattutto dopo la distribuzione. Tale attività dovrebbe riguardare tutte le componenti di un sistema di IA, compresi i dati, i modelli preaddrestrati, gli ambienti e il comportamento del sistema nel suo complesso e dovrebbe essere studiata e realizzata da un gruppo di persone il più eterogeneo possibile. Si dovrebbero sviluppare diverse metriche che racchiudano le categorie testate secondo diverse prospettive. È possibile valutare l'opportunità di impiegare diversi "red team" fidati che eseguano attacchi deliberati al sistema per trovarne le vulnerabilità e programmi di "bug bounty" che incentivino soggetti esterni a rilevare e segnalare in modo responsabile gli errori e i difetti del sistema. Infine, occorre garantire che gli esiti o le azioni siano coerenti con i risultati dei processi precedenti, confrontandoli con strategie predefinite per accertarsi che queste ultime non siano disattese.

- *Indicatori di qualità del servizio*

(102) È possibile definire adeguati indicatori di qualità del servizio che servano da base di riferimento per capire se i sistemi di IA sono stati testati e sviluppati tenendo conto delle questioni relative alla sicurezza e all'incolumità. Tali indicatori potrebbero comprendere sia misure per valutare le prove e l'addestramento degli algoritmi che le tradizionali metriche software per la valutazione di funzionalità, prestazione, utilizzabilità, affidabilità,

sicurezza, e manutenibilità.

2. Metodi non tecnici

(103) Questa sezione descrive una serie di metodi non tecnici che possono svolgere un ruolo importante nel garantire e mantenere un'IA affidabile. Anche questi ultimi dovrebbero essere valutati su **base continuativa**.

- *Regolamentazione*

(104) Come accennato in precedenza, la normativa a sostegno dell'affidabilità dell'IA esiste già, si pensi ad esempio alla normativa sulla sicurezza dei prodotti e ai quadri in materia di responsabilità. Questo aspetto sarà affrontato nelle raccomandazioni sugli investimenti e la politica in relazione all'IA, il nostro secondo documento, nella misura in cui riteniamo che potrebbe essere necessario rivedere o adattare la normativa vigente o introdurre nuove norme, sia come strumento di salvaguardia che di sostegno.

- *Codici di condotta*

(105) Le organizzazioni e i portatori di interessi possono sottoscrivere gli orientamenti e adattare la loro carta sulla responsabilità d'impresa, gli indicatori chiave di prestazione, i loro codici di condotta o i documenti di politica interna aggiungendo l'impegno per un'IA affidabile. Un'organizzazione che si occupa di un sistema di IA può, più in generale, non solo documentare le proprie intenzioni ma anche avallarle stabilendo livelli di auspicabilità di determinati valori, come i diritti fondamentali, la trasparenza e la prevenzione dei danni.

- *Normalizzazione*

(106) Le norme, ad esempio per la progettazione, la produzione e le pratiche commerciali, possono fungere da sistema di gestione della qualità per gli utenti dell'IA, i consumatori, le organizzazioni, gli istituti di ricerca e i governi, consentendo loro di riconoscere e incoraggiare una condotta etica tramite le loro decisioni di acquisto. Oltre alle norme convenzionali, esistono approcci di coregolamentazione: sistemi di accreditamento, codici deontologici professionali o norme per una progettazione conforme ai diritti fondamentali. Alcuni esempi attuali sono rappresentati dalle norme ISO o dalla serie di norme IEEE P7000, ma in futuro potrebbe essere opportuno prevedere una sorta di etichetta "IA affidabile", che confermi, facendo riferimento a norme tecniche specifiche, che il sistema risponde, ad esempio, ai requisiti di sicurezza, robustezza tecnica e spiegabilità.

- *Certificazione*

(107) Poiché non ci si può aspettare che tutti siano in grado di comprendere appieno il funzionamento e gli effetti dei sistemi di IA, occorre valutare la possibilità di ricorrere a organizzazioni che possano certificare per il pubblico generale che un sistema di IA è trasparente, responsabile ed equo[53]. Tali certificazioni applicherebbero norme elaborate per diversi campi di applicazione e diverse tecniche di IA, opportunamente allineate alle norme industriali e sociali dei vari contesti. La certificazione non può tuttavia mai sostituire la responsabilità e dovrebbe essere quindi integrata da quadri di accountability, tra cui clausole di esclusione della responsabilità nonché meccanismi correttivi e di riesame.[54]

- *Accountability tramite quadri di governance*

(108) Le organizzazioni dovrebbero istituire quadri di governance, sia interni che esterni, garantendo l'accountabillity per le dimensioni etiche delle decisioni associate allo sviluppo, alla distribuzione e all'utilizzo dell'IA. Ciò può includere, ad esempio, la nomina di una persona responsabile delle questioni etiche relative all'IA o di un

[53] Come sostenuto, ad esempio, dall'iniziativa IEEE Ethically Aligned Design Initiative: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Per ulteriori informazioni sulle limitazioni della certificazione, si veda: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

comitato/panel etico interno/esterno che abbia il compito, tra gli altri, di prestare sorveglianza e fornire consulenza. Come indicato in precedenza, anche le specifiche e/o gli organismi di certificazione possono essere utili a tal fine. Si dovrebbero garantire canali di comunicazione con l'industria e/o gruppi di controllo pubblico, condividendo le migliori pratiche, discutendo i dilemmi o segnalando le problematiche emergenti relative a preoccupazioni di natura etica. Tali meccanismi possono integrare la sorveglianza giuridica (ad esempio sotto forma di nomina di un responsabile della protezione dei dati o di misure equivalenti imposte dalla legislazione in materia di protezione dei dati), ma non sostituirla.

- *Istruzione e sensibilizzazione per promuovere una mentalità etica*

(109) Un'IA affidabile incoraggia la partecipazione informata di tutti i portatori di interessi. La comunicazione, l'istruzione e la formazione svolgono un ruolo importante, sia nel rendere ampiamente noti i possibili effetti dei sistemi di IA, sia per informare le persone che possono contribuire a plasmare lo sviluppo sociale. Tutti i portatori di interessi sono implicati, ad esempio coloro che si occupano della realizzazione dei prodotti (i progettisti e gli sviluppatori), gli utenti (imprese o singoli individui) e altri gruppi coinvolti (coloro che non possono acquistare o utilizzare un sistema di IA ma le cui decisioni sono adottate da tali sistemi, e la società in generale). L'alfabetizzazione di base sull'IA dovrebbe essere incoraggiata in tutta la società, ma il prerequisito per istruire il pubblico risiede nel garantire le abilità adeguate e la formazione di esperti di etica in questo ambito.

- *Partecipazione dei portatori di interessi e dialogo sociale*

(110) I vantaggi dell'IA sono molti e l'Europa deve garantire che siano alla portata di tutti. A tal fine occorre una discussione aperta e il coinvolgimento delle parti sociali, dei portatori di interessi e del pubblico in generale. Molte organizzazioni si affidano già a panel di portatori di interessi per discutere l'uso dei sistemi di IA e l'analisi dei dati. Tali panel sono composti da vari membri, quali giuristi, tecnici, esperti di etica, rappresentanti dei consumatori e lavoratori. Sollecitare la partecipazione e il dialogo sull'uso e sugli effetti dei sistemi di IA, oltre a contribuire alla valutazione dei risultati e degli approcci, può essere particolarmente utile in casi complessi.

- *Diversità e team di progettazione inclusivi*

(111) La diversità e l'inclusione sono elementi essenziali nello sviluppo di sistemi di IA da utilizzare nel mondo reale. Poiché i sistemi di IA svolgono svariati compiti autonomamente, è essenziale che i team che progettano, sviluppano, provano, eseguono la manutenzione, distribuiscono e/o acquistano questi sistemi riflettano la diversità degli utenti e della società in generale. In tal modo si contribuisce all'obiettività e alla considerazione di prospettive, esigenze e obiettivi diversi. Idealmente, tali team sono diversificati sia in termini di genere, cultura, età, che in termini di esperienze professionali e competenze.

Indicazioni chiave tratte dal capitolo II

- ✓ Garantire che l'intero ciclo di vita del sistema di IA soddisfi i requisiti per un'IA affidabile: 1) intervento e sorveglianza umani, 2) robustezza tecnica e sicurezza, 3) riservatezza e governance dei dati, 4) trasparenza, 5) diversità, non discriminazione ed equità, 6) benessere ambientale e sociale e 7) accountability.
- ✓ Prendere in considerazione metodi tecnici e non tecnici per garantire l'attuazione di tali requisiti.
- ✓ Favorire la ricerca e l'innovazione a sostegno della valutazione dei sistemi di IA e agevolare il rispetto dei requisiti; diffondere i risultati e le domande aperte al grande pubblico e formare sistematicamente una nuova generazione di esperti in etica dell'IA.
- ✓ Informare in modo chiaro e proattivo i portatori di interessi in merito alle capacità e ai limiti del sistema di IA, consentendo di creare aspettative realistiche, e in merito ai modi in cui i requisiti sono attuati. Essere trasparenti circa il fatto che si sta interagendo con un sistema di IA.

- ✓ Agevolare la tracciabilità e la verificabilità dei sistemi di IA, in particolare in contesti o situazioni critiche.
- ✓ Coinvolgere i portatori di interessi durante l'intero ciclo di vita del sistema di IA. Promuovere la formazione e l'istruzione affinché tutti i portatori di interessi siano formati e informati in merito all'IA affidabile.
- ✓ Essere consapevoli che vi potrebbero essere tensioni fondamentali tra i diversi principi e i diversi requisiti. Individuare, valutare, documentare e comunicare costantemente le soluzioni di compromesso.

III. Capitolo III - Valutazione di un'IA affidabile

(112) Il presente capitolo fornisce una **lista di controllo per la valutazione dell'affidabilità dell'IA** non esaustiva (versione pilota) per **rendere operativa un'IA affidabile**, sulla base dei requisiti fondamentali enunciati nel capitolo II. Tale lista di controllo è valida in particolare per i sistemi di IA che interagiscono direttamente con gli utenti, ed è principalmente destinata a sviluppatori e distributori di sistemi di IA (sviluppati internamente o acquisiti da terzi). La lista di controllo non concerne la prima componente dell'IA affidabile (legalità dell'IA). Il fatto di essersi attenuti a tale lista di controllo non è una prova di conformità giuridica, né tale lista va intesa come un orientamento per garantire la conformità alle leggi vigenti. Data la specificità applicativa dei sistemi di IA, la lista di controllo dovrà essere adattata ai casi d'uso e ai contesti specifici in cui operano i sistemi. Il presente capitolo contiene inoltre una raccomandazione generale sulle modalità di attuazione della lista di controllo per la valutazione dell'affidabilità dell'IA tramite una struttura di governance che include sia il livello operativo che quello gestionale.

(113) La lista di controllo e la struttura di governance saranno sviluppate in stretta collaborazione con i portatori di interessi del settore pubblico e privato. Il processo sarà condotto sotto forma di processo pilota, il che consentirà di raccogliere numerosi riscontri derivanti da due processi paralleli:

a. un processo qualitativo che garantisce la rappresentatività, a cui parteciperà una ristretta selezione di imprese, organizzazioni e istituzioni (appartenenti a diversi settori e di diverse dimensioni) che sperimenteranno la lista di controllo e la struttura di governance nella pratica e forniranno un riscontro approfondito;

b. un processo quantitativo, a cui potranno partecipare tutti i portatori di interessi per sperimentare la lista di controllo e fornire un riscontro tramite una consultazione aperta.

(114) Dopo la fase pilota inseriremo i risultati provenienti dai riscontri nella lista di controllo e all'inizio del 2020 prepareremo una versione riveduta. L'obiettivo è quello di ottenere un quadro di riferimento che possa essere utilizzato orizzontalmente per tutte le applicazioni e che funga quindi da base per garantire un'IA affidabile in tutti i settori. Una volta costituito tale nucleo fondamentale, si potrà eventualmente sviluppare un quadro settoriale o specifico per le applicazioni.

Governance

(115) Le imprese, le organizzazioni e le istituzioni che desiderano valutare come attuare la lista di controllo per la valutazione dell'affidabilità dell'IA nella loro organizzazione possono farlo includendo il processo di valutazione nei meccanismi di governance esistenti o attuando nuovi processi. La scelta dipenderà dalla struttura interna dell'organizzazione, dalle sue dimensioni e dalle risorse disponibili.

(116) Le ricerche[55] dimostrano che per ottenere un cambiamento è necessaria la massima attenzione da parte della dirigenza. È inoltre dimostrato che in un'impresa, in un'organizzazione o un'istituzione il coinvolgimento di tutti i portatori di interessi favorisce l'accettazione e avalla l'importanza

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

dell'introduzione di qualsiasi nuovo processo (tecnologico o meno)[56]. Raccomandiamo pertanto di adottare un processo che preveda il coinvolgimento sia del livello operativo sia dell'alta dirigenza.

Livello	Ruoli pertinenti (a seconda dell'organizzazione)
Dirigenza e consiglio di amministrazione	L'alta dirigenza discute e valuta lo sviluppo, la distribuzione o l'acquisto del sistema di IA e si occupa della procedura di escalation per la valutazione di tutte le innovazioni e gli utilizzi dell'IA, quando vengono rilevati elementi critici. Coinvolge in tutto il processo le persone interessate dall'eventuale introduzione di sistemi di IA (ad esempio i lavoratori) e i loro rappresentanti attraverso procedure di informazione, consultazione e partecipazione.
Dipartimento giuridico/di conformità/di responsabilità aziendale	Il dipartimento di responsabilità aziendale controlla l'utilizzo della lista di controllo e la sua necessaria evoluzione per far fronte ai cambiamenti tecnologici o normativi. Aggiorna le norme o le politiche interne relative ai sistemi di IA e garantisce che l'utilizzo di tali sistemi sia conforme al quadro giuridico e normativo vigente e ai valori dell'organizzazione.
Sviluppo di prodotti e servizi o equivalente	Il dipartimento per lo sviluppo di prodotti e servizi utilizza la lista di controllo per valutare prodotti e servizi basati sull'IA e registra tutti i risultati. Tali risultati sono discussi a livello della dirigenza, che approva in via definitiva le applicazioni nuove o riviste basate sull'IA.
Assicurazione della qualità	Il dipartimento di assicurazione della qualità (o equivalente) assicura e controlla i risultati della lista di controllo e segnala eventuali problemi a un livello gerarchico superiore (procedura di escalation) se il risultato non è soddisfacente o se vengono rilevati risultati imprevisti.
Risorse umane	Il dipartimento risorse umane assicura il giusto mix di competenze e la diversità dei profili degli sviluppatori dei sistemi di IA. Garantisce che all'interno dell'organizzazione venga fornito un adeguato livello di formazione sull'IA affidabile.
Acquisti	Il dipartimento acquisti assicura che il processo di acquisto di prodotti o servizi basati sull'IA preveda anche il controllo dell'affidabilità dell'IA.
Operazioni ordinarie	Gli sviluppatori e i responsabili di progetto includono la lista di controllo nelle loro attività quotidiane e documentano i risultati e gli esiti della valutazione.

Utilizzo della lista di controllo per la valutazione dell'affidabilità dell'IA

(117) Quando si utilizza la lista di controllo si raccomanda di prestare attenzione non solo agli aspetti problematici, ma anche alle domande che non trovano (facilmente) risposta. Un possibile problema potrebbe essere la mancanza di abilità e competenze differenziate nel team che sviluppa e testa il sistema di IA, e quindi potrebbe essere necessario coinvolgere altri portatori di interessi interni o esterni all'organizzazione. È vivamente raccomandato di registrare tutti i risultati sia in termini tecnici che

[56] Cfr. ad esempio A. Bryson, E. Barth e H. Dale-Olsen, "*The Effects of Organisational change on worker well-being and the moderating role of trade unions*", ILRReview, 66(4), luglio 2013; Jirjahn, U. and Smith, S.C. (2006) "*What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations*", 45(4), 650–680; Michie, J. and Sheehan, M. (2003) "*Labour market deregulation, 'flexibility' and innovation*", Cambridge Journal of Economics, 27(1), 123–143.

gestionali, assicurandosi che la soluzione dei problemi possa essere compresa a tutti i livelli della struttura di governance.

(118) La presente lista di controllo ha lo scopo di orientare gli operatori del settore dell'IA nello sviluppo, nella distribuzione e nell'utilizzo di un'IA affidabile. La valutazione dovrebbe essere adattata proporzionalmente al caso d'uso specifico. Durante la fase pilota potrebbero emergere ambiti sensibili: in tal caso la necessità di ulteriori specifiche sarà valutata nella fase successiva. La presente lista di controllo non fornisce risposte concrete per affrontare le problematiche emerse, ma incoraggia a riflettere su ciò che può contribuire a garantire l'affidabilità dei sistemi di IA e sulle possibili misure da adottare al riguardo.

Relazione con il diritto vigente e i processi esistenti

(119) È inoltre importante che coloro che partecipano allo sviluppo, alla distribuzione e all'utilizzo dell'IA riconoscano che esistono varie normative che impongono particolari processi e vietano determinati risultati e possono sovrapporsi e coincidere con alcune delle misure elencate nella lista di controllo. Ad esempio, la legislazione sulla protezione dei dati stabilisce una serie di obblighi giuridici che devono essere rispettati da coloro che si occupano della raccolta e del trattamento dei dati personali. Ebbene, poiché anche l'IA affidabile richiede il trattamento etico dei dati, le procedure interne e le politiche volte a garantire il rispetto della legislazione sulla protezione dei dati potrebbero anche contribuire ad agevolare il trattamento etico dei dati e possono quindi integrare i processi giuridici esistenti. La conformità alla lista di controllo *non* prova, tuttavia, la conformità giuridica, né tale lista va intesa come un orientamento per garantire la conformità alle leggi vigenti. Lo scopo della lista di controllo è piuttosto quello di porre una serie di domande specifiche a coloro che intendono garantire che il loro approccio allo sviluppo o alla distribuzione dell'IA è orientato verso un'IA affidabile, e tenta di garantire tale affidabilità.

(120) Analogamente, molti operatori del settore dell'IA dispongono già di strumenti di valutazione e di processi di sviluppo software per garantire la conformità anche a norme non giuridiche. La valutazione che segue non deve necessariamente essere effettuata come esercizio a sé stante, ma può essere integrata nelle pratiche già esistenti.

LISTA DI CONTROLLO PER LA VALUTAZIONE DELL'AFFIDABILITÀ DELL'IA (VERSIONE PILOTA)

1. Intervento e sorveglianza umani

Diritti fondamentali

- ✓ Nei casi d'uso che possono avere potenziali effetti negativi sui diritti fondamentali, è stata effettuata una valutazione d'impatto su tali diritti? Sono stati individuati e documentati i possibili compromessi realizzati tra i diversi principi e diritti?
- ✓ Il sistema di IA interagisce con il processo decisionale degli utenti finali umani (ad esempio azioni raccomandate o decisioni da prendere, presentazione di opzioni)?
  - In questi casi, sussiste il rischio che il sistema di IA influenzi l'autonomia umana interferendo in modo non intenzionale con il processo decisionale dell'utente finale?
  - È stato valutato se il sistema di IA debba comunicare agli utenti che una decisione, un contenuto, un consiglio o un risultato sono frutto di una decisione algoritmica?
  - Nel caso in cui il sistema di IA disponga di un chat bot o di un sistema conversazionale, gli utenti finali umani sono messi a conoscenza del fatto che stanno interagendo con un agente non

umano?

Intervento umano

- ✓ Nel caso in cui il sistema di IA sia implementato in processi lavorativi, è stato valutato se la distribuzione delle mansioni tra il sistema di IA e i lavoratori umani consente interazioni significative e un controllo e una sorveglianza umani adeguati?
  - Il sistema di IA migliora o aumenta le capacità umane?
  - Sono state adottate misure di salvaguardia per prevenire l'eccessiva fiducia nel sistema di IA o l'eccessiva dipendenza da esso nei processi di lavoro?

Sorveglianza umana

- ✓ È stato valutato quale sarebbe il livello appropriato di controllo umano per il sistema di IA e il caso d'uso in questione?
  - È possibile eventualmente descrivere il livello di controllo o di coinvolgimento umano? Chi è l'attore del "controllo umano" e quali sono i momenti o gli strumenti dell'intervento umano?
  - Sono stati messi in atto meccanismi e misure per garantire tali possibili controllo o sorveglianza umani o per garantire che le decisioni siano prese sotto la responsabilità generale di esseri umani?
  - Sono state adottate misure per rendere possibili verifiche e risolvere le problematiche relative alla gestione dell'autonomia dell'IA?
- ✓ Nel caso di un sistema di IA o di un caso d'uso autonomo o basato sull'autoapprendimento, sono stati messi in atto meccanismi più specifici di controllo e supervisione?
  - Che tipo di meccanismi di rilevamento e di risposta sono stati stabiliti per valutare l'eventualità di esiti avversi?
  - È stato previsto un "pulsante di arresto" o una procedura per interrompere in sicurezza un'operazione, se necessario? Questa procedura interrompe l'intero processo o una sua parte o delega il controllo a un essere umano?

2. <u>Robustezza tecnica e sicurezza</u>

Resilienza agli attacchi e sicurezza

- ✓ Sono state valutate le possibili forme di attacco a cui il sistema di IA potrebbe essere vulnerabile?
  - In particolare, sono stati presi in considerazione i diversi tipi di vulnerabilità e la loro natura, ad esempio l'inquinamento dei dati, le infrastrutture fisiche, gli attacchi informatici?
- ✓ Sono state messe in atto misure o sistemi per garantire l'integrità e la resilienza del sistema di IA contro possibili attacchi?
- ✓ È stato valutato il comportamento del sistema in situazioni e ambienti imprevisti?

- ✓ È stato valutato se e in che misura il sistema possa avere un duplice uso? In caso affermativo, sono state adottate misure preventive adeguate (ad esempio, non pubblicare le ricerche o non distribuire il sistema)?

Piano di emergenza e sicurezza generale

- ✓ È stato accertato che il sistema abbia un piano di emergenza adeguato nell'eventualità di *adversarial attack* o di altre situazioni impreviste (ad esempio, procedure di commutazione tecnica o la richiesta di intervento di un operatore umano prima di procedere)?
- ✓ È stato preso in considerazione il livello di rischio posto dal sistema di IA in questo caso d'uso specifico?
  - Sono stati messo in atto processi per misurare e valutare i rischi e la sicurezza?
  - Sono state fornite le informazioni necessarie in caso di rischio per l'integrità fisica umana?
  - È stata presa in considerazione la sottoscrizione di un'assicurazione contro eventuali danni provocati dal sistema di IA?
  - Sono stati identificati i rischi potenziali per la sicurezza posti da (altri) prevedibili usi della tecnologia, compreso l'uso improprio accidentale o malevolo? Esiste un piano per mitigare o gestire questi rischi?
- ✓ È stato valutato se vi sono probabilità che il sistema di IA possa arrecare danni o nuocere agli utenti o a terze parti? In caso affermativo, si è provveduto a valutare la probabilità, i possibili danni, i soggetti colpiti e la gravità?
  - Nel caso in cui vi sia il rischio che il sistema di IA provochi danni, sono state prese in considerazione le norme in materia di responsabilità e di protezione dei consumatori e come se ne è tenuto conto?
  - È stato preso in considerazione il possibile impatto ambientale o il rischio per la salute animale?
  - Nell'analisi dei rischi si è valutato se i problemi di sicurezza o di rete (ad esempio i rischi di sicurezza informatica) comportino rischi per la sicurezza o arrechino danni a causa di comportamenti non intenzionali del sistema di IA?
- ✓ Sono stati valutati i probabili effetti di un guasto del sistema di IA che comporti risultati errati, l'indisponibilità del sistema o risultati socialmente inaccettabili (ad es. pratiche discriminatorie)?
  - Nell'ambito degli scenari delineati sopra, sono stati definiti livelli minimi e regole in base ai quali attivare piani alternativi/di emergenza?
  - Sono stati definiti e testati i piani di emergenza?

Precisione

- ✓ È stato valutato quale livello e quale definizione di precisione sarebbero necessari nel contesto del sistema di IA e del caso d'uso?
  - È stato valutato come misurare e assicurare la precisione?
  - Sono state messe in atto misure per garantire che i dati utilizzati siano completi e aggiornati?

- Sono state messe in atto misure per valutare la necessità di dati aggiuntivi, ad esempio per migliorare la precisione o per eliminare le distorsioni?

✓ È stato valutato il danno che si verificherebbe se il sistema di IA facesse previsioni imprecise?

✓ Sono state previste modalità per misurare se il sistema effettua un numero inaccettabile di previsioni imprecise?

✓ Se il sistema effettua previsioni imprecise, è stata prevista una serie di misure per risolvere il problema?

Affidabilità e riproducibilità

✓ È stata messa in atto una strategia per monitorare e testare se il sistema di IA rispetta gli obiettivi, le finalità e le applicazioni previste?

- È stato verificato se è necessario prendere in considerazione contesti specifici o condizioni particolari per garantire la riproducibilità?
- Sono stati messi in atto processi o metodi di verifica per misurare e garantire diversi aspetti di affidabilità e riproducibilità?
- Sono stati messi in atto processi per descrivere quando, in date situazioni, un sistema di IA commette errori?
- Questi processi di prova e verifica dell'affidabilità dei sistemi di IA sono stati chiaramente documentati e resi operativi?

  Sono stati messi in atto meccanismi o metodi di comunicazione per garantire agli utenti (finali) che il sistema di IA è affidabile?

3. <u>Riservatezza e governance dei dati</u>

Rispetto della riservatezza e protezione dei dati

✓ Sono stati stabiliti meccanismi che, secondo i casi d'uso, consentano a terzi di segnalare problemi di riservatezza o di protezione dei dati riguardanti i processi del sistema di IA per la raccolta dati (sia per l'addestramento che per il funzionamento) e il loro trattamento?

✓ Sono stati valutati il tipo e la portata dei dati all'interno dei set (ad esempio se contengono dati personali)?

✓ Sono state prese in considerazione modalità per sviluppare il sistema di IA o addestrare il modello senza utilizzare dati personali o potenzialmente sensibili o utilizzandoli il meno possibile?

✓ Sono stati previsti meccanismi di notifica e controllo dei dati personali a seconda dei casi d'uso (come il consenso valido e la possibilità di revoca, se del caso)?

✓ Sono state adottate misure per aumentare la riservatezza, come la crittografia, l'anonimizzazione e l'aggregazione?

✓ Nel caso in cui esista un responsabile della protezione dei dati personali, è stato coinvolto nel processo fin dall'inizio?

Qualità e integrità dei dati

- ✓ Per quanto riguarda la gestione ordinaria e la governance dei dati, il sistema è conforme alle eventuali norme pertinenti (ad esempio ISO, IEEE) o ai protocolli più diffusi?
- ✓ Sono stati stabiliti meccanismi di sorveglianza relativi alla raccolta, all'archiviazione, all'elaborazione e all'utilizzo dei dati?
- ✓ È stato valutato in che misura si ha il controllo della qualità delle fonti di dati esterne utilizzate?
- ✓ Sono stati messi in atto processi per garantire la qualità e l'integrità dei dati? Sono stati presi in considerazione altri processi? Come viene verificato che i set di dati non vengano compromessi o violati?

Accesso ai dati

- ✓ Quali protocolli, processi e procedure sono stati seguiti per gestire e garantire una corretta governance dei dati?
  - ▪ È stato valutato chi può accedere ai dati degli utenti e in quali circostanze?
  - ▪ È stato accertato che queste persone siano qualificate, siano autorizzate ad accedere ai dati e che abbiano le competenze necessarie per comprendere la politica di protezione dei dati nei dettagli?
  - ▪ È stato previsto un meccanismo di sorveglianza che registri sia chi ha avuto accesso ai dati sia i tempi, i luoghi, le modalità e le finalità di accesso?

4. <u>Trasparenza</u>

Tracciabilità

- ✓ Sono state messe in atto misure che garantiscono la tracciabilità? A tal fine potrebbe essere necessario documentare gli elementi elencati di seguito.
  - ➢ Metodi utilizzati per la progettazione e lo sviluppo del sistema algoritmico:
    - o nel caso di un sistema di IA basato su regole, dovrebbe essere documentato il metodo di programmazione o il modo in cui il modello è stato costruito;
    - o nel caso di un sistema di IA basato sull'apprendimento, dovrebbe essere documentato il metodo di addestramento dell'algoritmo, quali dati di input sono stati raccolti e selezionati e come ciò è avvenuto.
  - ➢ Metodi utilizzati per testare e convalidare il sistema algoritmico:
    - o nel caso di un sistema di IA basato su regole, dovrebbero essere documentati gli scenari o i casi utilizzati per testare e convalidare;
    - o nel caso di un modello basato sull'apprendimento, dovrebbero essere documentate le informazioni sui dati utilizzati per testare e convalidare.
  - ➢ Risultati del sistema algoritmico:
    - o dovrebbero essere documentati i risultati dell'algoritmo o le decisioni adottate in base ad esso, come pure altre possibili decisioni che potrebbero derivare da casi diversi (ad

esempio per altri sottogruppi di utenti).

Spiegabilità

- ✓ È stato valutato fino a che punto le decisioni prese dal sistema di IA e i loro risultati sono comprensibili?
- ✓ È stato accertato che la spiegazione dei motivi per cui un sistema ha adottato una certa decisione che ha prodotto determinati risultati possa essere resa comprensibile per gli utenti che la desiderano?
- ✓ È stato valutato in quale misura le decisioni del sistema influenzano i processi decisionali dell'organizzazione?
- ✓ È stato valutato il motivo per cui un particolare sistema è stato distribuito in un dato settore?
- ✓ È stato valutato il modello di business relativo al sistema (ad esempio in quale modo crea valore per l'organizzazione)?
- ✓ Il sistema di IA è stato progettato tenendo presente fin dall'inizio l'interpretabilità?
  - Sono state effettuate ricerche sul modello più semplice e più interpretabile possibile per l'applicazione in questione e si è provato ad usarlo?
  - È stata valutata la possibilità di analizzare i dati di addestramento e di test? È possibile modificarli e aggiornarli nel corso del tempo?
  - È stato valutato se vi è la possibilità di esaminare l'interpretabilità dopo aver addestrato e sviluppato il modello o se è possibile accedere al flusso di lavoro interno del modello?

Comunicazione

- ✓ È stato comunicato agli utenti (finali), tramite una clausola di esclusione della responsabilità o altri mezzi, che stanno interagendo con un sistema di IA e non con un altro essere umano? Il sistema di IA è stato etichettato come tale?
- ✓ Sono stati messi in atto meccanismi per informare gli utenti in merito alle ragioni e ai criteri che determinano i risultati del sistema di IA?
  - Tali informazioni sono comunicate in modo chiaro e comprensibile agli utenti destinatari?
  - Sono stati stabiliti processi che prendono in considerazione i riscontri degli utenti e li utilizzano per adattare il sistema?
  - Sono stati comunicati anche i rischi potenziali o percepiti, come le distorsioni (*bias*)?
  - A seconda del caso d'uso, sono state prese in considerazione anche la comunicazione e la trasparenza nei confronti di terzi o del pubblico in generale?
- ✓ È stato precisato qual è lo scopo del sistema di IA e chi o cosa può trarre vantaggio dal prodotto/servizio?
  - Gli scenari d'uso del prodotto sono descritti dettagliatamente e comunicati chiaramente, eventualmente utilizzando anche forme di comunicazione alternative, per garantire che l'informazione sia comprensibile e adeguata per gli utenti destinatari?
  - A seconda del caso d'uso, si è riflettuto sulla psicologia umana e su potenziali limiti, come il

rischio di confusione, il pregiudizio di conferma (*confirmation bias*) o la fatica cognitiva?

- ✓ Le caratteristiche, i limiti e le potenziali carenze del sistema di IA sono stati chiaramente comunicati:
  - nel caso dello sviluppo, a chiunque lo distribuisca in un prodotto o servizio?
  - nel caso della distribuzione, all'utente finale o al consumatore?

5. <u>Diversità, non discriminazione ed equità</u>

Evitare distorsioni inique

- ✓ È stata prevista una strategia o una serie di procedure per evitare di creare o rafforzare distorsioni inique (*unfair bias*) nel sistema di IA, sia per quanto riguarda l'uso dei dati di input che per la progettazione dell'algoritmo?
  - Sono state valutate e riconosciute le possibili limitazioni derivanti dalla composizione dei set di dati utilizzati?
  - Sono state prese in considerazione la diversità e la rappresentatività degli utenti nei dati? Si è provveduto a effettuare test per popolazioni specifiche o casi d'uso problematici?
  - Sono state effettuate ricerche relative agli strumenti tecnici disponibili per migliorare la comprensione dei dati, del modello e delle prestazioni? Tali strumenti sono utilizzati?
  - Sono stati messi in atto processi per testare e monitorare eventuali distorsioni durante le fasi di sviluppo, distribuzione e utilizzo del sistema?
- ✓ A seconda del caso d'uso, è stato previsto un meccanismo che permetta a terzi di segnalare problemi di distorsione, discriminazione o scarsa prestazione del sistema di IA?
  - Sono stati presi in considerazione misure e metodi di comunicazione chiari in merito a come e a chi esporre tali problematiche?
  - Sono stati presi in considerazione, oltre agli utenti (finali), anche terze parti che potrebbero essere indirettamente interessate dal sistema di IA?
- ✓ È stato valutato se esiste la possibilità che si verifichi una variabilità delle decisioni nelle stesse condizioni?
  - In caso affermativo, sono state valutate le possibili cause?
  - In caso di variabilità, è stato stabilito un meccanismo di misurazione o di valutazione del potenziale impatto di tale variabilità sui diritti fondamentali?
- ✓ È stata prevista un'adeguata definizione operativa di "equità" da applicare alla progettazione di sistemi di IA?
  - La definizione è di uso comune? Sono state prese in considerazione altre definizioni prima di scegliere quella utilizzata?
  - Sono stati previsti un'analisi quantitativa o parametri per misurare e testare la definizione di equità utilizzata?

- Sono stati stabiliti meccanismi per garantire l'equità dei sistemi di IA? Sono stati presi in considerazione altri possibili meccanismi?

Accessibilità e progettazione universale

✓ È stato accertato che il sistema di IA si adegui a una gamma estesa di preferenze e abilità individuali?

- È stato valutato se il sistema di IA è utilizzabile da persone con bisogni speciali, disabilità o a rischio di esclusione? In che modo questo aspetto è stato inserito nella progettazione del sistema e come viene verificato?
- È stato accertato che le informazioni sul sistema di IA siano accessibili anche agli utenti di tecnologie assistive?
- Questa categoria di persone è stata consultata durante la fase di sviluppo del sistema di IA?

✓ Si è tenuto conto degli effetti del sistema di IA sugli utenti potenziali?

- Il team coinvolto nella realizzazione del sistema di IA è rappresentativo degli utenti destinatari? È rappresentativo della popolazione in senso lato, considerando anche altri gruppi che potrebbero essere marginalmente interessati?
- È stato valutato vi sono se persone o gruppi che potrebbero essere interessati in maniera sproporzionata da conseguenze negative?
- Sono pervenuti riscontri da altri team o gruppi con esperienze e background diversi?

Partecipazione dei portatori di interessi

✓ È stato preso in considerazione un meccanismo per far partecipare allo sviluppo e all'utilizzo del sistema di IA diversi portatori di interessi?

✓ L'introduzione del sistema di IA nell'organizzazione è stata preparata informando e coinvolgendo in anticipo i lavoratori interessati e i loro rappresentanti?

6. <u>Benessere sociale e ambientale</u>

IA sostenibile e rispettosa dell'ambiente

✓ Sono stati messi in atto meccanismi per misurare l'impatto ambientale dello sviluppo, della distribuzione e dell'utilizzo del sistema di IA (ad esempio, la quantità e il tipo di energia utilizzata dal centro dati, ecc.)

✓ Sono state previste misure per ridurre l'impatto ambientale del ciclo di vita del sistema di IA?

Impatto sociale

✓ Qualora il sistema di IA interagisca direttamente con gli esseri umani:

- È stato valutato se il sistema di IA incoraggia gli esseri umani a sviluppare attaccamento ed empatia verso il sistema?
- È stato accertato che il sistema di IA segnali chiaramente che la sua interazione sociale è simulata e che non ha capacità né di "capire" né di "provare sentimenti"?

- ✓ È stato accertato che l'impatto sociale del sistema di IA sia ben compreso? Ad esempio, è stato valutato se esiste un rischio di perdita di posti di lavoro o di dequalificazione della forza lavoro? Quali misure sono state adottate per contrastare tali rischi?

Società e democrazia

- ✓ È stato valutato l'impatto più generale sulla società dell'uso del sistema di IA al di là del singolo utente (finale), come ad esempio i portatori di interessi indirettamente interessati?

7. <u>Accountability</u>

Verificabilità

- ✓ Sono stati messi in atto meccanismi che facilitano la verificabilità del sistema da parte di soggetti interni e/o indipendenti, come ad esempio la tracciabilità e la registrazione dei processi e dei risultati del sistema di IA?

Riduzione al minimo degli effetti negativi e loro segnalazione

- ✓ È stata effettuata una valutazione del rischio o dell'impatto del sistema di IA che tenga conto dei diversi portatori di interessi direttamente o indirettamente interessati?
- ✓ Sono stati messi in atto sistemi di formazione e istruzione per sviluppare pratiche di accountability?
  - Quali lavoratori o settori del team sono coinvolti? Tale processo va oltre la fase di sviluppo?
  - Nei corsi di formazione si insegna anche il possibile quadro giuridico applicabile al sistema di IA?
  - È stata presa in considerazione l'istituzione di un "comitato di revisione etica dell'IA" o un meccanismo analogo per discutere di accountability e pratiche etiche in generale, comprese le "zone grigie" potenzialmente poco chiare?
- ✓ Oltre alle iniziative interne o ai quadri per la supervisione etica e dell'accountability, esiste qualche tipo di orientamento esterno o sono stati istituiti anche processi di verifica?
- ✓ Sono stati messi in atto processi che permettano a terzi (ad esempio fornitori, consumatori, distributori/venditori) o ai lavoratori di segnalare potenziali vulnerabilità, rischi o distorsioni nel sistema/applicazione dell'IA?

Documentazione dei compromessi

- ✓ È stato stabilito un meccanismo per individuare gli interessi e i valori pertinenti chiamati in causa dal sistema di IA e i potenziali compromessi tra di essi?
- ✓ Con quale processo si adottano decisioni in merito a tali compromessi? È stata garantita la documentazione della decisione adottata in merito ai compromessi?

Capacità di ricorso

- ✓ È stata stabilita una serie adeguata di meccanismi di ricorso in caso di danni o effetti negativi?
- ✓ Soni stati messi in atto meccanismi per fornire informazioni sia agli utenti (finali) sia a terzi sulle possibilità di ricorso?

Invitiamo tutti i portatori di interessi a sperimentare nella pratica la presente lista di controllo e a fornire un riscontro sulla sua attuabilità, la sua completezza, la sua pertinenza per l'applicazione o il settore specifici dell'IA, come pure sulla sua sovrapposizione o complementarietà con i processi di conformità o di valutazione esistenti. In base a tali riscontri, all'inizio del 2020 verrà proposta alla Commissione una versione riveduta della lista di controllo.

Indicazioni chiave tratte dal capitolo III

- ✓ Adottare la **lista di controllo** per la valutazione dell'affidabilità dell'IA nelle fasi di sviluppo, distribuzione o utilizzo dell'IA e adattarla allo specifico caso d'uso del sistema.
- ✓ Tenere presente che tale lista di controllo **non sarà mai esaustiva**. Garantire che un'IA affidabile non sia una questione di caselle da spuntare, ma un processo continuo di individuazione dei requisiti, valutazione delle soluzioni e miglioramento dei risultati durante l'intero ciclo di vita del sistema di IA e di coinvolgimento dei portatori di interessi in tale processo.

C. ESEMPI DI OPPORTUNITÀ E SERIE PREOCCUPAZIONI DERIVANTI DELL'IA

(121) Nella sezione seguente forniamo esempi di sviluppo e utilizzo dell'IA che dovrebbero essere incoraggiati ed esempi in cui lo sviluppo, la distribuzione o l'utilizzo dell'IA possono essere in contrasto con i nostri valori e destare specifiche preoccupazioni. Si deve trovare un equilibrio tra ciò che dovrebbe e ciò che può essere fatto con l'IA senza assolutamente trascurare ciò che non dovrebbe essere fatto con essa.

1. Esempi di opportunità di un'IA affidabile

(122) Un'IA affidabile può rappresentare una grande opportunità per contribuire a mitigare le pressanti sfide che la società si trova ad affrontare, ad esempio l'invecchiamento della popolazione, la crescente disuguaglianza sociale e l'inquinamento ambientale. Questo potenziale si riflette anche a livello mondiale, come nel caso degli obiettivi di sviluppo sostenibile delle Nazioni Unite[57]. La sezione seguente esamina come incoraggiare una strategia europea per l'IA che affronti alcune di queste sfide.

a. Azione per il clima e infrastrutture sostenibili

(123) Non vi è dubbio che la lotta ai cambiamenti climatici debba essere una priorità assoluta per i responsabili politici di tutto il mondo, ma la trasformazione digitale e un'IA affidabile offrono grandi potenzialità in termini di riduzione dell'impatto umano sull'ambiente e di contributo a un uso efficiente ed efficace dell'energia e delle risorse naturali[58]. Un'IA affidabile, ad esempio, può essere accoppiata ai Big Data per rilevare con maggiore precisione il fabbisogno energetico e favorire di conseguenza infrastrutture e consumi energetici più efficienti.[59]

(124) Nel settore del trasporto pubblico i sistemi di IA per i trasporti intelligenti[60] possono essere utilizzati per ridurre al minimo le code, ottimizzare i percorsi, consentire alle persone con deficit visivi di essere più

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Diversi progetti dell'UE, volti allo sviluppo di reti intelligenti e allo stoccaggio di energia, possono contribuire al successo della transizione energetica basata sulle tecnologie digitali, anche grazie a soluzioni basate sull'IA e ad altre soluzioni digitali. Per integrare il lavoro di questi singoli progetti, la Commissione ha lanciato l'iniziativa BRIDGE che consente ai progetti sulle reti intelligenti e sullo stoccaggio dell'energia nell'ambito di Orizzonte 2020 di creare una visione comune su questioni trasversali: https://www.h2020-bridge.eu/.

59 Cfr. ad esempio il progetto Encompass: http://www.encompass-project.eu/.

60 Nuove soluzioni basate sull'intelligenza artificiale aiutano a preparare le città al futuro della mobilità. Cfr. ad esempio il progetto Fabulos, finanziato dall'UE: https://fabulos.eu/.

indipendenti[61] e ottimizzare i motori ad alta efficienza energetica, contribuendo così agli sforzi di decarbonizzazione e alla riduzione dell'impatto ambientale, a favore di una società più verde. Attualmente nel mondo ogni 23 secondi un essere umano muore in un incidente d'auto[62]. I sistemi di IA potrebbero contribuire a ridurre significativamente il numero di vittime, ad esempio migliorando i tempi di reazione e il rispetto delle regole.[63]

b. Salute e benessere

(125) Le tecnologie di IA affidabili possono essere utilizzate (e già lo sono) per rendere le terapie più intelligenti e mirate e contribuire a prevenire malattie potenzialmente letali[64]. I medici e gli operatori sanitari possono potenzialmente eseguire un'analisi più accurata e dettagliata dei complessi dati sanitari di un paziente, anche prima che si ammali, e prescrivere una terapia preventiva ad hoc[65]. Nel contesto dell'invecchiamento della popolazione europea, l'IA e la robotica possono essere strumenti preziosi per aiutare gli anziani o chi li assiste[66] e per monitorare le condizioni dei pazienti in tempo reale, salvando così delle vite umane.[67]

(126) Un'IA affidabile può fornire contributi anche su scala più ampia. Ad esempio, può esaminare e identificare le tendenze generali nel settore sanitario e terapeutico[68], contribuendo a diagnosticare precocemente le malattie, a sviluppare medicinali in modo più efficiente, a decidere terapie più mirate[69] e, in ultima analisi, a salvare un maggior numero di vite.

c. Istruzione di qualità e trasformazione digitale

(127) Visti i nuovi cambiamenti tecnologici, economici e ambientali, occorre che la società diventi più proattiva. I

---

61 Cfr. ad esempio il progetto PRO4VIP, nell'ambito della strategia europea Vision 2020 per combattere la cecità prevenibile, soprattutto a causa della vecchiaia. La mobilità e l'orientamento sono stati due settori prioritari del progetto.

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 Il progetto europeo UP-Drive, ad esempio, affronta problemi emergenti nel settore dei trasporti fornendo contributi che consentono l'automazione graduale dei veicoli e la loro reciproca collaborazione, contribuendo a un sistema di trasporto più sicuro, più inclusivo e più accessibile. https://up-drive.eu/.

64 Cfr. per esempio il progetto REVOLVER (*Repeated Evolution of Cancer*): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, o il progetto Murab per eseguire biopsie più precise e che mira a diagnosticare più velocemente il cancro e altre malattie: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Cfr. ad esempio il progetto Live INCITE: www.karolinska.se/en/live-incite. Questo consorzio di acquirenti pubblici di servizi sanitari chiede all'industria di sviluppare l'IA e altre soluzioni TIC che consentano di intervenire sullo stile di vita nel processo perioperatorio. L'obiettivo riguarda nuove soluzioni innovative di eHealth che possono influenzare i pazienti in modo personalizzato intervenendo sul loro stile di vita, sia prima che dopo l'intervento chirurgico, per ottimizzare il risultato sanitario/terapeutico.

66 Il progetto CARESSES, finanziato dall'UE, si occupa di robot per l'assistenza agli anziani, concentrandosi sulla sensibilità culturale: il robot adatta il modo di agire e di parlare alla cultura e alle abitudini dell'anziano che assiste: http://caressesrobot.org/en/project/. http://caressesrobot.org/it/project/. Cfr. anche l'applicazione di IA Alfred, un assistente virtuale che aiuta gli anziani a rimanere attivi: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Inoltre, il progetto EMPATTICS (*EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems*) ricercherà e definirà come gli operatori sanitari e i pazienti utilizzano le tecnologie TIC, compresi i sistemi di IA, per pianificare interventi con i pazienti e per monitorare l'evoluzione del loro stato fisico e psichico: www.empattics.eu.

67 Cfr. ad esempio MyHealth Avatar (www.myhealthavatar.eu), che offre una rappresentazione digitale dello stato di salute di un paziente. Il progetto di ricerca ha lanciato un'applicazione e una piattaforma online che raccoglie e dà accesso alle informazioni digitali sullo stato di salute a lungo termine e assume la forma di un "avatar sanitario" che accompagna l'utente per tutta la vita. MyHealthAvatar prevede anche il rischio di ictus, di diabete, di malattie cardiovascolari e ipertensione.

68 Cfr. ad esempio il progetto ENRICHME (www.enrichme.eu), che propone soluzioni al problema della progressiva diminuzione delle capacità cognitive nella popolazione che invecchia. Gli anziani possono rimanere indipendenti e attivi più a lungo tramite una piattaforma integrata per l'ambiente domestico assistito (*Ambient Assisted Living* - AAL) e un androide dedicato al monitoraggio e all'interazione a lungo termine.

69 Cfr. ad esempio l'uso dell'IA da parte di Sophia Genetics, che sfrutta l'inferenza statistica, il riconoscimento di modelli e l'apprendimento automatico per massimizzare il valore dei dati genomici e radiomici: https://www.sophiagenetics.com/home.html.

governi, i leader dell'industria, gli istituti di istruzione e i sindacati hanno la responsabilità di portare i cittadini nella nuova era digitale garantendo che abbiano le abilità necessarie per occupare i futuri posti di lavoro. Le tecnologie dell'IA affidabile possono aiutare a prevedere con maggiore precisione quali posti di lavoro e quali professioni saranno trasformati dalla tecnologia, quali nuovi ruoli saranno creati e quali abilità saranno necessarie. I governi, i sindacati e l'industria potrebbero usufruire di tali informazioni per pianificare la (ri)qualificazione dei lavoratori, mentre i cittadini a rischio di esubero potrebbero essere inseriti in un percorso di sviluppo per rivestire nuovi ruoli.

(128) L'IA può essere inoltre un utile strumento per combattere le disuguaglianze nel campo dell'istruzione e creare programmi di insegnamento personalizzati e adattabili per aiutare ad acquisire nuove qualifiche, abilità e competenze in base alle capacità di apprendimento di ciascuno.[70] Potrebbe aumentare sia la velocità di apprendimento sia la qualità dell'istruzione dalla scuola primaria all'università.

2. Esempi di serie preoccupazioni destate dall'IA

(129) La violazione di una delle componenti dell'IA affidabile dà luogo a serie preoccupazioni. Molte delle preoccupazioni elencate di seguito ricadono già nel campo di applicazione di norme giuridiche vigenti, che sono vincolanti e devono quindi essere rispettate. Tuttavia, anche quando l'osservanza delle norme giuridiche sia dimostrata, resta il fatto che tali norme potrebbero non risolvere le preoccupazioni etiche che possono insorgere. Poiché il nostro modo d'intendere l'adeguatezza delle norme e dei principi etici invariabilmente evolve e muta nel tempo, la serie non esaustiva di esempi di preoccupazioni riportata di seguito potrà in futuro essere ridotta, ampliata, modificata o aggiornata.

a. Identificazione e tracciamento degli individui con l'IA

(130) L'IA permette a organismi pubblici e privati di identificare in modo sempre più efficiente le singole persone. Esempi degni di nota di una tecnologia scalabile di identificazione tramite IA sono il riconoscimento facciale e altri metodi di identificazione involontaria che utilizzano dati biometrici (ad esempio il rilevamento della menzogna, la valutazione della personalità in base a micro espressioni e il rilevamento automatico della voce). Identificare una persona talvolta è un atto auspicabile e rispettoso dei principi etici (ad esempio nei casi di frode, riciclaggio di denaro o finanziamento del terrorismo). L'identificazione automatica tuttavia desta enormi preoccupazioni di natura sia giuridica che etica, in quanto può avere effetti non previsti sotto molti aspetti a livello psicologico e socioculturale. Per salvaguardare l'autonomia dei cittadini europei è necessario ricorrere alle tecniche di controllo tramite l'IA in modo proporzionato. Definire chiaramente se, quando e come l'IA può essere utilizzata per l'identificazione automatica degli individui e differenziare tra l'identificazione di un individuo e la sua tracciatura e localizzazione, e tra sorveglianza mirata e sorveglianza di massa, sarà fondamentale per ottenere un'IA affidabile. L'applicazione di tali tecnologie deve essere chiaramente motivata dal diritto vigente[71]. Se la base giuridica di tale attività è rappresentata dal "consenso"[72], devono essere sviluppati mezzi pratici che permettano di dare un consenso eloquente e verificato ad essere identificati automaticamente da un sistema di IA o da tecnologie equivalenti. Ciò vale anche per l'utilizzo di dati personali "anonimi" che possono essere ripersonalizzati.

b. Sistemi di IA nascosti

---

70 Cfr. ad esempio il progetto MaTHiSiS, volto a fornire una soluzione per l'apprendimento basato sullo stato affettivo in un ambiente di apprendimento confortevole che prevede dispositivi tecnologici e algoritmi di alto livello: (http://mathisis-project.eu/it). Cfr. anche la piattaforma IBM Watson Classroom o la piattaforma Century Tech.

71 A questo proposito, può essere ricordato l'articolo 6 del regolamento generale sulla protezione dei dati, secondo il quale, tra l'altro, il trattamento dei dati è lecito solo se ha una base giuridica valida.

72 Come dimostrato dagli attuali meccanismi che permettono di dare il consenso informato in Internet, i consumatori di solito acconsentono senza prestare la debita attenzione. È difficile quindi definire tali meccanismi efficaci.

(131) Gli esseri umani dovrebbero sempre sapere se stanno interagendo direttamente con un altro essere umano o con una macchina e spetta agli operatori del settore dell'IA la responsabilità di comunicarlo in modo affidabile. Gli operatori del settore dell'IA dovrebbero pertanto garantire (ad esempio tramite clausole di esclusione della responsabilità chiare e trasparenti) che le persone siano consapevoli del fatto che stanno interagendo con un sistema di IA o che possano chiedere informazioni in merito e approvare tale interazione. Va notato che esistono casi limite che complicano la questione (ad esempio una voce umana filtrata da un sistema di IA). Occorre tenere presente che la confusione tra esseri umani e macchine potrebbe avere molteplici conseguenze quali attaccamento, influenza o svilimento dell'essere umano[73]. Lo sviluppo di androidi[74] dovrebbe pertanto essere oggetto di un'attenta valutazione etica.

c. Valutazione per punteggio dei cittadini tramite l'IA in violazione dei diritti fondamentali

(132) Nelle società la libertà e l'autonomia di tutti i cittadini dovrebbero essere tutelate. Qualsiasi forma di valutazione per punteggio delle persone può comportare la perdita di tale autonomia e compromettere il principio di non discriminazione. La valutazione per punteggio dovrebbe essere utilizzata solo se esiste una chiara giustificazione e se le misure sono proporzionate ed eque. La valutazione normativa dei cittadini per punteggio (valutazione generale della "personalità morale" o dell'"integrità etica"), in *tutti gli* aspetti e su larga scala, effettuata da autorità pubbliche o soggetti privati compromette questi valori, soprattutto se utilizzata in violazione dei diritti fondamentali, in modo sproporzionato e senza uno scopo legittimo descritto e comunicato.

(133) Al giorno d'oggi la valutazione delle persone per punteggio si utilizza spesso, su larga o piccola scala, per scopi puramente descrittivi e specifici del settore (ad esempio, nei sistemi scolastici, nell'ambito dell'e-learning e per le patenti di guida), ma anche in questi casi più limitati, dovrebbe essere disponibile una procedura pienamente trasparente, che fornisca informazioni sul processo, sullo scopo e sulla metodologia di attribuzione del punteggio. Va notato che la trasparenza non può impedire la discriminazione né garantire l'equità e non è la panacea contro il problema della valutazione per punteggio. Idealmente, quando possibile, dovrebbero essere previste modalità per dissociarsi dal meccanismo di valutazione per punteggio, senza subire alcun pregiudizio, altrimenti dovrebbero essere previsti meccanismi per contestare e rettificare il punteggio. Ciò è particolarmente importante in situazioni in cui esiste un'asimmetria di potere tra le parti. Le possibilità di dissociazione dovrebbero essere garantite nella progettazione della tecnologia laddove si debba garantire il rispetto dei diritti fondamentali e sia necessario per una società democratica.

d. Sistemi d'arma autonomi letali (LAWS)

(134) Attualmente, un numero imprecisato di paesi e industrie si dedica alla ricerca e allo sviluppo di sistemi d'arma autonomi letali, come missili capaci di selezionare gli obiettivi o macchine ad apprendimento automatico con abilità cognitive che consentono di decidere chi, quando e dove combattere senza l'intervento umano. Questa situazione pone interrogativi etici fondamentali, per esempio la possibilità di una corsa incontrollabile agli armamenti a un livello storicamente senza precedenti e la creazione di contesti militari in cui il controllo umano è quasi del tutto assente e i rischi di malfunzionamento non sono presi in considerazione. Il Parlamento europeo ha chiesto l'elaborazione urgente di una posizione comune giuridicamente vincolante che affronti questioni etiche e giuridiche fondamentali relative al controllo e alla supervisione da parte dell'uomo, all'accountability e all'attuazione del diritto internazionale in materia di diritti umani, del diritto internazionale umanitario e delle strategie militari.[75] Ricordando che l'Unione europea si prefigge di promuovere la pace come sancito dall'articolo 3 del trattato sull'Unione europea, sosteniamo la risoluzione del Parlamento europeo del 12 settembre 2018 e tutti gli sforzi correlati in materia di sistemi d'arma autonomi letali.

---

73 Madary & Metzinger (2016), "*Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology",* Frontiers in Robotics and IA, 3(3).

74 Questo vale anche per gli avatar guidati dall'IA.

75 Risoluzione 2018/2752(RSP) del Parlamento europeo.

e. Potenziali preoccupazioni a lungo termine

(135) Lo sviluppo dell'IA è ancora specifico per settore e richiede scienziati e ingegneri umani adeguatamente formati che ne definiscano con precisione gli obiettivi. Procedendo per deduzione, con un orizzonte temporale più lungo, si può tuttavia ipotizzare l'emergere di gravi preoccupazioni a lungo termine[76]. Secondo l'approccio basato sul rischio tali preoccupazioni dovrebbero essere tenute in considerazione alla luce di eventuali incognite sconosciute e dei cosiddetti "cigni neri"[77]. La natura dirompente di tali preoccupazioni assieme all'attuale incertezza sui relativi sviluppi, impone una valutazione regolare di questi temi.

D. CONCLUSIONI

(136) Il presente documento contiene gli orientamenti etici sull'IA elaborati dal gruppo di esperti ad alto livello sull'intelligenza artificiale.

(137) Riconosciamo che i sistemi di IA hanno già avuto e continueranno ad avere effetti positivi sia a livello commerciale che sociale, ma ci preme ugualmente garantire che i rischi e gli altri effetti negativi a cui queste tecnologie sono associate siano gestiti in modo adeguato e proporzionato all'applicazione dell'IA in questione. L'IA è una tecnologia trasformativa e di rottura e la sua evoluzione negli ultimi anni è stata agevolata dalla disponibilità di enormi quantità di dati digitali, dai grandi progressi tecnologici nella potenza di calcolo e nella capacità di memorizzazione e da una importante innovazione scientifica e ingegneristica nei metodi e negli strumenti di IA. I sistemi di IA continueranno ad avere effetti sulla società e sui cittadini che non possiamo ancora immaginare.

(138) In tale contesto è importante realizzare sistemi di IA che siano degni di fiducia, poiché gli esseri umani potranno sfruttarne pienamente e con sicurezza i vantaggi solo se la tecnologia, i processi e le persone a monte si riveleranno affidabili. Nel redigere i presenti orientamenti l'IA affidabile è stata quindi la nostra ambizione fondamentale.

(139) Un'IA affidabile si basa su tre componenti: 1) legalità, l'IA deve ottemperare a tutte le leggi e ai regolamenti applicabili; 2) eticità, l'IA deve assicurare l'adesione a principi e valori etici e 3) robustezza, dal punto di vista tecnico e sociale poiché, anche con le migliori intenzioni, i sistemi di IA possono causare danni non intenzionali. Ciascuna componente è necessaria ma non sufficiente per realizzare un'IA affidabile. Idealmente le tre componenti operano armonicamente e si sovrappongono; In caso di tensioni occorre adoperarsi per risolverle.

(140) Nel capitolo I sono stati trattati i diritti fondamentali e un corpus corrispondente di principi etici essenziali nel contesto dell'IA. Nel capitolo II sono stati elencati sette requisiti chiave che i sistemi di IA devono soddisfare per realizzare un'IA affidabile e sono stati proposti metodi tecnici e non tecnici che possono contribuire alla loro attuazione. Infine, nel capitolo III è stata proposta una lista di controllo per la valutazione dell'affidabilità dell'IA che può aiutare a rendere operativi i sette requisiti. Nella sezione finale sono stati illustrati esempi di opportunità vantaggiose e di serie preoccupazioni che i sistemi di IA suscitano e sui quali speriamo di incoraggiare ulteriori discussioni.

(141) L'Europa gode di un vantaggio esclusivo, che deriva dal suo impegno a porre il cittadino al centro delle proprie attività. Tale impegno è iscritto nel DNA stesso dell'Unione europea attraverso i trattati su cui si fonda. Il presente documento rientra in una visione che promuove un'IA affidabile la quale, a nostro avviso, dovrebbe costituire il presupposto su cui l'Europa può sviluppare la propria leadership nei sistemi di IA innovativi e

[76] Alcuni ritengono che l'intelligenza generale artificiale, la coscienza artificiale, gli agenti morali artificiali, la superintelligenza o l'IA trasformativa possano essere esempi di tali preoccupazioni a lungo termine (attualmente inesistenti), mentre molti altri ritengono che siano irrealistiche.

[77] Un "cigno nero" è un evento estremamente raro ma dirompente, talmente raro che potrebbe non essere stato mai osservato, per cui la probabilità che si verifichi può essere stimata solo con elevata incertezza.

all'avanguardia. Questa visione ambiziosa contribuirà a garantire la prosperità dei cittadini europei, sia a livello individuale che collettivo. Il nostro obiettivo è quello di creare una cultura dell'"IA affidabile per l'Europa", che permetta a tutti di sfruttarne i vantaggi in un modo che garantisca il rispetto dei nostri valori fondamentali: i diritti fondamentali, la democrazia e lo Stato di diritto.

GLOSSARIO

(142) Il presente glossario fa parte degli orientamenti ed è inteso ad agevolare la comprensione dei termini usati nel documento.

Intelligenza artificiale o sistemi di IA

(143) Sistemi software (ed eventualmente hardware) progettati dall'uomo[78] che, dato un obiettivo complesso, agiscono nella dimensione fisica o digitale percependo il proprio ambiente attraverso l'acquisizione di dati, interpretando i dati strutturati o non strutturati raccolti, ragionando sulla conoscenza o elaborando le informazioni derivate da questi dati e decidendo le migliori azioni da intraprendere per raggiungere l'obiettivo dato. I sistemi di IA possono usare regole simboliche o apprendere un modello numerico, e possono anche adattare il loro comportamento analizzando gli effetti che le loro azioni precedenti hanno avuto sull'ambiente.

(144) Come disciplina scientifica, l'IA comprende diversi approcci e diverse tecniche, come l'apprendimento automatico (di cui l'apprendimento profondo e l'apprendimento per rinforzo sono esempi specifici), il ragionamento meccanico (che include la pianificazione, la programmazione, la rappresentazione delle conoscenze e il ragionamento, la ricerca e l'ottimizzazione) e la robotica (che comprende il controllo, la percezione, i sensori e gli attuatori e l'integrazione di tutte le altre tecniche nei sistemi ciberfisici).

(145) Il gruppo di esperti ad alto livello sull'intelligenza artificiale ha elaborato un documento distinto che definisce più in dettaglio i *sistemi di IA*, utilizzato ai fini del presente documento e pubblicato in parallelo, intitolato "Una definizione di IA: principali capacità e discipline scientifiche".

Operatori del settore dell'IA

(146) Individui o organizzazioni che sviluppano (compresa la ricerca, la progettazione o la fornitura di dati), distribuiscono (compresa l'implementazione) o utilizzano sistemi di IA, a esclusione di coloro che li utilizzano in qualità di utenti finali o di consumatori.

Ciclo di vita del sistema di IA

(147) Ciclo di vita che comprende le seguenti fasi: sviluppo (compresa la ricerca, la progettazione, la fornitura di dati e alcune sperimentazioni), distribuzione (compresa l'implementazione) e utilizzo.

Verificabilità

(148) Possibilità di un sistema di IA di essere soggetto alla valutazione dei suoi algoritmi, dati e processi di progettazione. È uno dei 7 requisiti che un sistema di IA affidabile deve soddisfare. Ciò non implica necessariamente che le informazioni sui modelli di business e sulla proprietà intellettuale del sistema di IA debbano essere sempre apertamente disponibili. Garantire la tracciabilità e i meccanismi di registrazione sin dalle prime fasi di progettazione del sistema di IA può contribuire alla sua verificabilità.

Distorsione (*bias*)

(149) È la tendenza al pregiudizio nei confronti di una persona, un oggetto o una posizione. Nei sistemi di IA la distorsione si può verificare in molti modi. Ad esempio, nei sistemi di IA basati sui dati, come quelli prodotti tramite l'apprendimento automatico, le distorsioni nella raccolta dei dati e nell'addestramento possono dar luogo a un sistema di IA che presenta distorsioni. Nell'IA basata sulla logica, come i sistemi basati su regole, possono verificarsi distorsioni dovute al modo in cui un ingegnere della conoscenza interpreta le regole che si applicano in un particolare ambiente. Le distorsioni si possono verificare anche a causa dell'apprendimento online e dell'adattamento tramite l'interazione e possono derivare dalla personalizzazione in base alla quale gli utenti ricevono raccomandazioni o informazioni adattate ai loro gusti. Non si riferiscono necessariamente a pregiudizi umani o alla raccolta di dati su iniziativa umana. Possono derivare, ad esempio, dall'utilizzo del

[78] Gli esseri umani progettano direttamente i sistemi di IA, ma possono anche utilizzare tecniche di IA per ottimizzarne la progettazione.

sistema in contesti limitati che non consentono la generalizzazione ad altri contesti. Le distorsioni possono essere benevole o malevole, intenzionali o non intenzionali. In alcuni casi, le distorsioni possono portare a risultati discriminatori e/o iniqui, che nel presente documento sono definiti distorsioni inique.

Etica

(150) L'etica è una disciplina accademica e una branca della filosofia. In termini generali, l'etica si occupa di domande come "Cos'è una buona azione?", "Qual è il valore di una vita umana?", "Cos'è la giustizia?" o "Cos'è la felicità?". A livello accademico esistono quattro principali campi di ricerca etica: i) metaetica, riguarda principalmente la connotazione e la denotazione dell'enunciato normativo, e la questione di come stabilire i loro valori di verità(se ve ne sono), ii) etica normativa, i mezzi pratici per determinare la moralità di una condotta esaminando le norme alla base di un'azione giusta o ingiusta e attribuendo un valore alle azioni specifiche, iii) etica descrittiva, riguarda l'indagine empirica relativa ai comportamenti e alle convinzioni morali degli individui, e iv) etica applicata, riguarda ciò che si deve (o si può) fare in determinate circostanze (spesso storicamente nuove) o in un particolare settore d'azione (spesso storicamente senza precedenti). L'etica applicata si occupa di situazioni di vita reale, in cui le decisioni devono essere prese rapidamente e spesso con limitata razionalità. L'etica dell'IA è generalmente considerata un esempio di etica applicata e si concentra sulle questioni normative sollevate dalla progettazione, dallo sviluppo, dall'implementazione e dall'utilizzo dell'IA.

(151) Nell'ambito del discorso relativo all'etica, spesso ricorrono i termini "morale" ed "etico". Il termine "morale" si riferisce a modelli concreti e fattuali di comportamento, ai costumi e alle convenzioni propri di specifiche culture, gruppi o individui in un determinato momento. Il termine "etico" si riferisce al giudizio estimativo di tali azioni e comportamenti concreti a partire da una prospettiva sistematica e accademica.

Eticità dell'IA

(152) Nel presente documento, per eticità dell'IA si intende lo sviluppo, la distribuzione e l'utilizzo dell'IA tali da garantire il rispetto delle norme etiche, compresi i diritti fondamentali in quanto diritti morali speciali, dei principi etici e dei valori fondamentali connessi. È la seconda delle tre componenti essenziali e necessarie di un'IA affidabile.

IA antropocentrica

(153) L'approccio antropocentrico all'IA è volto a garantire che i valori umani rivestano un ruolo centrale nelle modalità di sviluppo, distribuzione, utilizzo e monitoraggio dei sistemi di IA, garantendo il rispetto dei diritti fondamentali, tra cui quelli sanciti nei trattati dell'Unione europea e nella Carta dei diritti fondamentali dell'Unione europea, accomunati dal riferimento a un fondamento condiviso radicato nel rispetto della dignità umana, nei quali l'essere umano gode di uno status morale unico e inalienabile. Ciò implica anche il rispetto dell'ambiente naturale e di altri esseri viventi che fanno parte dell'ecosistema umano e un approccio sostenibile che consenta alle generazioni future di prosperare.

Red Teaming

(154) Pratica in cui un "red team" o un gruppo indipendente sfida un'organizzazione a migliorare la propria efficacia assumendo il ruolo o punto di vista di avversario. Tale pratica è impiegata soprattutto nell'ambito della sicurezza per contribuire a individuare e risolvere le potenziali vulnerabilità.

Riproducibilità

(155) La riproducibilità indica se un esperimento di IA mostra lo stesso comportamento quando ripetuto nelle stesse condizioni.

Robustezza dell'IA

(156) La robustezza di un sistema di IA comprende sia la robustezza tecnica (adeguata ad un dato contesto, come

l'ambito di applicazione o la fase del ciclo di vita) sia la robustezza in termini sociali (ossia la garanzia che il sistema di IA tenga debitamente conto del contesto e dell'ambiente in cui esso opera). Questa componente è essenziale per garantire che, anche con le migliori intenzioni, non si producano danni non intenzionali. La robustezza è la terza delle tre componenti essenziali e necessarie di un'IA affidabile.

Portatori di interessi

(157) Tutti coloro che svolgono attività di ricerca, sviluppano, progettano, distribuiscono o utilizzano l'IA e tutti coloro che sono (direttamente o indirettamente) interessati dall'IA, compresi, a titolo esemplificativo, aziende, organizzazioni, ricercatori, servizi pubblici, istituzioni, organizzazioni della società civile, governi, autorità di regolamentazione, parti sociali, soggetti privati, cittadini, lavoratori e consumatori.

Tracciabilità

(158) In riferimento a un sistema di IA è la capacità di tenere traccia dei dati del sistema, dei processi di sviluppo e di distribuzione, generalmente tramite l'identificazione registrata documentata.

Fiducia

(159) In letteratura è presente la seguente definizione. La fiducia è intesa come: 1) un insieme di credenze relative a benevolenza, competenza, integrità e prevedibilità (credenze legate alla fiducia); 2) la volontà di una parte di dipendere da un'altra in una situazione di rischio (intenzione di fidarsi); oppure 3) una combinazione di questi due elementi.[79] La "fiducia" di solito non è una proprietà attribuita alle macchine, ma in questo documento si intende sottolineare l'importanza di potersi fidare non solo del fatto che i sistemi di IA ottemperano a norme giuridiche, aderiscono a principi etici e sono robusti, ma anche del fatto che tale fiducia può essere riconosciuta a tutte le persone e ai processi coinvolti nel ciclo di vita del sistema di IA.

IA affidabile

(160) Un'IA affidabile si basa su tre componenti: 1) legalità, l'IA deve ottemperare a tutte le leggi e ai regolamenti applicabili, 2) eticità, l'IA deve dimostrare il rispetto di principi e valori etici e assicurarvi l'adesione e 3) robustezza, dal punto di vista tecnico e sociale poiché, anche con le migliori intenzioni, i sistemi di IA possono causare danni non intenzionali. Un'IA affidabile non comporta solo l'affidabilità del sistema di IA stesso, ma anche l'affidabilità di tutti i processi e degli attori che fanno parte del ciclo di vita del sistema.

Persone e gruppi vulnerabili

(161) Non esiste una definizione giuridica comunemente accettata o ampiamente condivisa di persone vulnerabili, in quanto si tratta di una categoria estremamente eterogenea. Spesso ciò che rende una persona o un gruppo vulnerabile è un elemento specifico del contesto: si può trattare di eventi temporanei della vita (come l'infanzia o la malattia), fattori di mercato (come l'asimmetria informativa o il potere di mercato), fattori economici (come la povertà), fattori legati alla propria identità (come il genere, la religione o la cultura) o altri fattori. L'articolo 21 della Carta dei diritti fondamentali dell'Unione europea vieta la discriminazione fondata sui seguenti motivi, che possono costituire un punto di riferimento: sesso, razza, colore della pelle, origine etnica o sociale, caratteristiche genetiche, lingua, religione o convinzioni personali, opinioni politiche o di qualsiasi altra natura, appartenenza a una minoranza nazionale, patrimonio, nascita, disabilità, età e orientamento sessuale. Altri articoli della Carta dei diritti fondamentali dell'Unione europea riguardano i diritti di gruppi specifici, oltre a quelli sopra elencati. Tale elenco non è esaustivo e può subire variazioni nel tempo. Un gruppo vulnerabile è un gruppo di persone che condividono una o più caratteristiche di vulnerabilità.

---

79 Siau, K., Wang, W. (2018), "*Building Trust in Artificial Intelligence, Machine Learning, and Robotics*", CUTTER BUSINESS TECHNOLOGY JOURNAL (31), S. 47–53.

Il presente documento è stato elaborato dai membri del gruppo di esperti ad alto livello sull'IA

elencati di seguito in ordine alfabetico

Pekka Ala-Pietilä, presidente del gruppo di esperti ad alto livello sull'IA
IA Finlandia, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Correlatore
Zalando
Mária Bieliková
Università slovacca di tecnologia di Bratislava
Cecilia Bonefeld-Dahl – Correlatrice
DigitalEurope
Yann Bonnet
ANSSI

Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems e Università della Sorbona
Mark Coeckelbergh
Università di Vienna
Virginia Dignum – Correlatrice
Università di Umeå
Luciano Floridi
Università di Oxford
Jean-Francois Gagné – Correlatore
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agenzia per i diritti fondamentali
Sami Haddadin
Munich School of Robotics and MI
Gry Hasselbalch
The thinkdotank DataEthics e Università di Copenaghen
Fredrik Heintz
Università di Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Università di Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
Università tecnica di Vienna
Robert Kroplewski
Avvocato e consulente del governo polacco
Elisabeth Ling
RELX

Pierre Lucas
Orgalim – Industrie tecnologiche europee

Ieva Martinkenaite
Telenor
Thomas Metzinger – Correlatore
JGU Mainz e Associazione europea delle università
Catelijne Muller
ALLAI Netherlands e CESE
Markus Noga
SAP
Barry O'Sullivan, vicepresidente del gruppo di esperti ad alto livello sull'IA
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – Correlatore
Università di Liegi
Christoph Peylo
Bosch

Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
Facoltà del Collegio d'Europa e CEPS
Francesca Rossi
IBM
Cristina San José
Federazione bancaria europea
George Sharkov
Digital SME Alliance
Philipp Slusallek
Centro di ricerca tedesco sull'intelligenza artificiale (DFKI)
Françoise Soulié Fogelman
Consulente di IA
Saskia Steinacker – Correlatrice
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Correlatrice
Università tecnica di Delft
Thiébaut Weber
CES
Cecile Wendling
AXA
Karen Yeung – Correlatrice
Università di Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe e Karen Yeung hanno assolto alla funzione di relatori per il presente documento.

Pekka Ala-Pietilä presiede il gruppo di esperti ad alto livello sull'IA. Il vicepresidente del gruppo di esperti ad alto livello sull'IA,Barry O'Sullivan, ha coordinato l'elaborazione del secondo documento. Nozha Boujemaa, vicepresidente fino al 1º febbraio 2019, ha coordinato l'elaborazione del primo documento e ha contribuito alla stesura del presente documento.

Nathalie Smuha ha fornito assistenza editoriale.