Valentina Cavosi

GOVERNARE L'INTELLIGENZA ARTIFICIALE

SPUNTI PER LA PROGETTAZIONE DI SISTEMI DI INTELLIGENZA ARTIFICIALE LEGALI, ETICI E ROBUSTI

Introduzione di Federico Cabitza
Postfazione di Donatella Paschina

ReD OPEN FACTORY
A CENTER FOR RESPONSIBLE INNOVATION

Ledizioni
The Innovative LEDIpublishing Company

Quaderni della ReD OPEN FACTORY

Valentina Cavosi

GOVERNARE L'INTELLIGENZA ARTIFICIALE

SPUNTI PER LA PROGETTAZIONE DI SISTEMI DI INTELLIGENZA ARTIFICIALE LEGALI, ETICI E ROBUSTI

Introduzione di Federico Cabitza

Postfazione di Donatella Paschina

Ledizioni

2022 Ledizioni LediPublishing
Via Boselli 10, 20136 Milano
http://www.ledizioni.it
e-mail: info@ledizioni.it

Prima edizione Ledizioni: febbraio 2022

Valentina Cavosi, *Governare l'intelligenza artificiale. Spunti per la progettazione di sistemi di intelligenza artificiale legali, etici e robusti*

ISBN cartaceo 978-88-5526-622-2

In copertina: photo by Sander Weeteling on unsplash.com

Informazioni sul catalogo e sulle ristampe: www.ledizioni.it

Indice

Introduzione

Prof. Federico Cabitza
Università di Milano Bicocca
Comitato Scientifico ReD OPEN Factory

Assistiamo ad una crescente diffusione di sistemi digitali in grado di elaborare grandi quantità di dati, anche in formato non strutturato, e di applicare ad essi regole che nessun essere umano ha scritto direttamente, ma che sono state prodotte mediante tecniche e metodiche di "apprendimento automatico" (*machine learning*), per generare "contenuti, previsioni, raccomandazioni o decisioni che influenzano l'ambiente con cui tali sistemi interagiscono"[1], inclusi i suoi utenti. Per queste loro caratteristiche, sempre più frequentemente si assegna a queste tecnologie digitali l'etichetta di sistemi in grado di esibire o esprimere una certa "Intelligenza Artificiale", cioè la capacità di eseguire autonomamente compiti complessi che si pensa richiedano un certo grado (e un certo tipo) di intelligenza agli esseri umani per eseguirli correttamente.

Non è un caso che abbiamo introdotto questa locuzione – Intelligenza Artificiale – con cautela e in

[1] Proposta di regolamento del parlamento europeo e del consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) 2021/0106(COD). https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

maniera così perifrastica; infatti, temiamo che l'espressione stessa, in virtù della sua popolarità e forza mediatica (forse eccessive di questi tempi), possa distogliere l'attenzione da una semplice considerazione.

Quando si parla della applicazione dell'Intelligenza Artificiale in contesti organizzativi, non si fa riferimento ad altro che non sia l'automazione di operazioni che possono avere un impatto rilevante in processi aziendali dove siano elaborati molti dati, personali e non (i cosiddetti "big data" se vogliamo usare un'altra espressione altrettanto vaga e parimenti abusata), e in cui alcuni esseri umani sono chiamati a prendere decisioni che in molti casi riguardano altri esseri umani (ad esempio fornitori, clienti, impiegati, dirigenti), oppure che riguardano indirettamente un più ampio novero di possibili portatori di interesse.

I sistemi che chiamiamo Intelligenza Artificiale non riguardano perciò solo funzionalità particolarmente avanzate, sorprendenti o più tipiche di scenari fantascientifici, ma anche contesti aziendali e sociali che conosciamo e frequentiamo tutti i giorni.

In questo volume parleremo di come valutare questi sistemi: lo faremo partendo dall'estensione e diffusione di un lavoro che è stato oggetto di una tesi di laurea per il corso di laurea magistrale in Teoria e Tecnologia della Comunicazione dell'Università degli Studi di Milano-Bicocca, che ho personalmente seguito e che rappresenta un punto d'avvio per approfondimenti e implementazioni sul campo.

Riteniamo quindi opportuno riflettere sull'importanza dei processi di valutazione di qualsiasi tecnologia

che abbia un potenziale impatto sugli esseri umani, e di conseguenza anche – e soprattutto – sull'importanza di valutare i rischi connessi alle tecnologie informatiche che abilitano o potenziano processi amministrativi, analitici o decisionali al centro della catena del valore delle aziende.

Questi processi di valutazione della tecnologia informatica sono infatti necessari per comprendere diverse cose:

- se una tale tecnologia avrà un impatto sull'organizzazione che l'adotta per l'automazione, parziale o totale, di certi processi;
- se il suo impatto sarà positivo, cioè se i vantaggi e gli aspetti positivi saranno maggiori degli svantaggi che con maggiore o minore probabilità si concretizzeranno in seguito alla sua adozione; e, infine,
- se l'eventuale impatto positivo sarà valso lo sforzo, cioè se l'adozione di questa tecnologia in esame sarà 'costo-efficace' e il ritorno sull'investimento, non solo economico, sarà significativo.

Questi tre aspetti costituiscono il nucleo di qualsiasi attività di valutazione della tecnologia (*technology assessment*), a prescindere dai metodi e dalle tecniche messe in campo per produrla; infatti, il primo passo di qualsiasi attività consulenziale per valutare l'impatto di una tecnologia informatica è la convinzione consapevole che nessuno degli aspetti che abbiamo citato debba essere dato per scontato, neppure per le tecnologie più avanzate, come quelle a cui associamo l'etichetta di "Intelligenza Artificiale". Ad esempio, una tecnologia potrebbe sì portare benefici all'azienda che la adotta,

ma risultare insostenibile, sia economicamente che socialmente (e, perché no, anche dal punto di vista dei consumi energetici e dell'impronta di carbonio); oppure essa potrebbe essere addirittura dannosa, portando più svantaggi che vantaggi, o svantaggi più rilevanti; oppure, più semplicemente, una tecnologia digitale potrebbe essere inutile, nel senso che, a conti fatti, non ha alcun impatto sul lavoro di tutti i giorni. Quest'ultimo scenario, anche se sembra il più scandaloso e improbabile in molti circoli consulenziali, è invece molto più comune in ambito aziendale di quanto si creda: lo riteniamo infatti molto frequente anche nell'attuale panorama delle applicazioni di Intelligenza Artificiale progettate per supportare gli ambiti professionali in cui se ne parla di più, quali la medicina, la gestione delle risorse umane, il giornalismo e la pratica legale.

Il lavoro che segue tratta ed esplora lo spazio di idee e concetti in cui si possono produrre e comprendere i risultati di una valutazione dell'impatto di sistemi di Intelligenza Artificiale, nella più ampia e concreta accezione menzionata sopra. Per questo motivo esso si rivolge a qualunque decisore, amministratore, o responsabile di processo che operi in contesti organizzativi, aziendali e istituzionali allo scopo di aiutarlo a comprendere alcuni termini chiave (come trasparenza, robustezza, affidabilità, o utilità) dal significato solo apparentemente comune, per scoprire la complessità e, purtroppo ambiguità, di certe definizioni; e per fornirgli un riferimento agile e accessibile per esercitare il governo responsabile e informato di processi decisionali informatizzati complessi, tra cui ovviamente

anche quelli che riguardano l'acquisto e adozione di un sistema software commercializzato sotto il nome di Intelligenza Artificiale (o certificato come tale, come in ambito sanitario). Infatti ogni sistema di Intelligenza Artificiale concepito e progettato per il contesto aziendale, non è tanto o solo uno strumento informatico come tanti altri di "burotica" e automazione di ufficio, al contrario è anche, e soprattutto, il motore e l'occasione per un più ampio cambiamento organizzativo che riguarda le modalità e priorità di produzione e gestione del dato – che in larga parte è personale e, come tale, richiede l'applicazione del regolamento del GDPR –, e le modalità in cui alcune importanti decisioni sono prese in azienda sulla base di quel dato.

Il perché occuparsi di questi temi dovrebbe essere chiaro. Le motivazioni che spingono verso crescenti livelli di informatizzazione di processi decisionali strategici ed operativi sono molteplici: la promessa di incrementi di efficienza (a parità di risultati o di risorse impiegate); una minore variabilità di criterio o esito (anche come conseguenza di una minore dipendenza dalla esperienza, arbitrio o pregiudizio del decisore umano); una maggiore efficacia (in termini o di minore tasso di errore o di maggiore qualità degli effetti della decisione); e infine, per limitarsi alle motivazioni principali, una maggiore ripetibilità e standardizzazione del processo, per garantire tanto una maggiore imparzialità per i soggetti coinvolti quanto una maggiore resilienza dell'azienda nei confronti di fenomeni di rotazione o avvicendamento delle sue risorse umane. Sono cose note; ma a queste considerazioni è giunto il

tempo di aggiungere la consapevolezza degli elementi di fragilità e rischio inerenti alle tecnologie più avanzate, quali la loro vulnerabilità ad attacchi dall'esterno a scopo estorsivo (cyberattacchi o *adversarial attacks*); la dipendenza da quantità di dati la cui qualità e rappresentatività, anche per la loro dimensione e le fonti di dati disponibili, è difficile da verificare e garantire; la loro capacità di esacerbare disuguaglianze o perpetrare ingiustizie nei confronti di minoranze etniche, di genere o, peggio, di accesso alle risorse (ad esempio, per età, istruzione, o capacità economiche). A questa consapevolezza si deve affiancare una nuova sensibilità per la valutazione dell'impatto di iniziative di informatizzazione della decisione nei termini della sostenibilità sociale (relativa alle trasformazioni dell'occupazione della forza lavoro) e umana (che riguarda il rischio che l'automazione impoverisca l'*expertise* delle risorse umane e i processi di acquisizione di nuove conoscenze, o ne causi una lenta ma progressiva perdita di competenze o deresponsabilizzazione).

A differenza dei proclami che si sentono spesso levare dall'industria e da alcuni operatori del mercato, l'era delle macchine intelligenti non è ancora arrivata; o almeno non si è ancora concretizzata l'età in cui le macchine sono in grado di sostituire o influenzare, con autorità e ingombrante autorevolezza, i processi decisionali più delicati che caratterizzano le nostre aziende e organizzazioni. Anziché essere una notizia deprimente o demotivante per il lettore che si accinge ad affrontare le prossime pagine, questa è un'ottima cosa: infatti siamo ancora in tempo per acquisire quella cul-

tura di innovazione, o sarebbe meglio dire, progresso, che ci permetterà di giudicare con equilibrio e competenza l'opportunità di digitalizzare alcuni processi e di capire, con tempestività e adeguatezza, quali siano le modalità migliori per farlo con trasparenza, equità, rispetto del principio di legalità, e responsabilità.

Ritengo che il trasferimento tecnologico di risultati selezionati che derivano da studi e ricerche condotte in ambito accademico sia un'operazione opportuna e necessaria, soprattutto in questa fase di grande rinnovamento degli strumenti digitali che si rendono a disposizione del lavoro di analisi e dei processi decisionali in ambito organizzativo: il volume si situa in questo filone di attività, che segue e che rafforza la collaborazione e sinergia tra il mio centro di ricerca e la ReD Open Factory.

Guida alla lettura

Lo spirito che anima il lavoro che segue è quello di rendere disponibili anche al lettore non esperto dei materiali per facilitare la conoscenza e l'approfondimento di un tema che sta tuttora alimentando discussioni e conversazioni nei contesti giuridici, governativi e industriali sia in Italia che all'estero: l'Intelligenza Artificiale e il suo uso.

Tale tema, in continua evoluzione e sviluppo, influenza anche l'iter normativo, tuttora in via di definizione; proprio per questo motivo il lavoro è utile per chi vuole comprendere i principi fondamentali e di conseguenza poter seguire e approfondire quanto sta accadendo a livello geopolitico.

L'introduzione del prof. Federico Cabitza, e la postfazione di Donatella Paschina, sono gli ulteriori contributi che introducono e allargano il contesto di riflessione sul tema di questo lavoro.

Infatti, al di là del percorso normativo e giuridico di una futura normativa sull'uso di tecniche di cosiddetta "Intelligenza Artificiale", è fondamentale che si abbia consapevolezza dell'impatto che queste evolute tecniche di automazione potranno avere sulle vite quotidiane di tutti noi.

Oggigiorno, inoltre, esse arricchiscono la trasformazione digitale in corso da anni, con la conseguenza che, in aggiunta ai principi di riservatezza e privacy dell'individuo, si aggiunge il tema del governo delle tecniche di decisioni che impattano sulla vita delle persone e

dell'ambiente circostante.

Quello che segue è perciò un contributo specialistico per orientarsi sul tema, il cui aggiornamento è in corso all'interno di ReD OPEN Factory, il "CENTER FOR RESPONSIBLE INNOVATION" per la governance della trasformazione digitale.

Buona lettura.

Per ulteriori approfondimenti: www.redopenfactory.com

1. Opportunità e sfide dell'Intelligenza Artificiale

I sistemi di Intelligenza Artificiale (IA), ovvero sistemi informatici che automatizzano compiti complessi che richiedono agli esseri umani capacità cognitive superiori, quali il riconoscimento di schemi e forme e il ragionamento logico-deduttivo, sono considerati tra le principali tecnologie che caratterizzano la quarta rivoluzione industriale e la distinguono dalle precedenti. Grazie alle recenti tecniche di apprendimento automatico, agli algoritmi sempre più sofisticati, alla concentrazione di sempre maggiori capacità di calcolo e alla crescente disponibilità di grandi quantità di dati (*big data*), sono stati raggiunti risultati impressionanti nei campi in cui l'IA si è potuta applicare, soprattutto quando a questi risultati si confrontano i medesimi conseguiti invece dagli esseri umani (in condizioni controllate e sperimentali). L'Intelligenza Artificiale è utilizzata in un sempre crescente numero di settori, tra cui quello della pubblica amministrazione (aiutando ad esempio a velocizzare i processi di assunzione, di assistenza domiciliare, di indennità di malattia o di disoccupazione), della sanità (supportando i medici nelle decisioni che riguardano la previsione, la diagnosi e la caratterizzazione di condizioni patologiche con accuratezza ed efficienza) e dell'agricoltura (aumentandone l'efficienza al fine di ridurre l'uso di acqua, fertilizzanti e pesticidi).

Sebbene i vantaggi dell'IA siano ampiamente riconosciuti sia per il settore pubblico sia per quello privato, il suo utilizzo solleva anche alcune preoccupazioni e timori, soprattutto dal punto di vista etico, legale e sociale, in termini di responsabilità, sicurezza, privacy, restrizioni alla libertà di espressione, pregiudizi ingiusti, violazioni della dignità umana o discriminazioni. Se non progettata correttamente, infatti, l'IA può portare a decisioni distorte nei procedimenti penali, nel diritto a determinati benefici sociali, nelle assunzioni o nei licenziamenti. L'Intelligenza Artificiale può anche compromettere gravemente il diritto alla privacy e protezione dei dati, ad esempio quando viene utilizzata per il riconoscimento facciale, il monitoraggio o la profilazione online degli individui; può anche essere utilizzata per creare video, audio o immagini falsi (*deep fake*), con conseguenti rischi finanziari e reputazionali. Si possono leggere molti casi reali, infatti, dove queste tecnologie comportano un forte rischio di uso improprio intenzionale o accidentale che minaccia i diritti fondamentali. Pensiamo ad esempio al noto caso del progetto della chatbot Tay di Microsoft: nato per raccontare barzellette, rispondere a domande oppure offrire commenti su fotografie inviate dagli utenti, ha imparato presto ad essere razzista (per esempio molti tweet di Tay hanno fatto riferimento a Hitler, negato l'Olocausto o sostenuto i piani di immigrazione di Trump). Oppure pensiamo al caso del sistema di reclutamento di Amazon, che ha imparato da solo a preferire i candidati maschi penalizzando tutti i curriculum che includevano la parola "donna" e declassando i diplo-

mati di due college femminili; all'algoritmo utilizzato per categorizzare in modo automatico le fotografie in Google Photo, che aveva iniziato a identificare le persone di colore come gorilla; oppure al filtro di FaceApp che faceva apparire le persone di colore come più bianche quando invece dovevano apparire semplicemente più belle. Joy Buolamwini e Timnit Gebru in uno studio[1] sull'accuratezza dei prodotti di classificazione di genere basati sull'IA e offerti da aziende quali IBM, Microsoft e Face ++, hanno dimostrato che questi prodotti riescono a classificare in modo più accurato gli uomini bianchi rispetto alle donne di colore. Questi tassi di errore in contesti di pubblica sicurezza possono portare persone innocenti ad essere segnalate: nel 2020 è accaduto che un sistema di riconoscimento facciale ha portato un uomo di colore all'arresto per un crimine non commesso.

Anche nel contesto della pubblica amministrazione emergono non poche preoccupazioni: come possiamo leggere nel rapporto del servizio di conoscenza della Commissione Europea[2], vari paesi dell'Olanda, per rilevare con più efficacia le frodi sul welfare, hanno utilizzato il sistema *"System Risk Indication"* (SyRI). Sviluppato dal Governo olandese, esso utilizza grandi

[1] Joy Buolamwini and Timnit Gebru, *Gender shades: Intersectional accuracy disparities in Commercial gender classification. Conference on fairness, accountability and transparency*, PMLR, 2018.

[2] Gianluca Misuraca and Colin Van Noordt, *AI Watch – Artificial Intelligence in public services*, Publications Office, 2020.

quantità di dati personali per rilevare rischi di frodi o uso improprio delle prestazioni sociali. A inizio 2020 la Corte internazionale di giustizia ha dichiarato che l'uso di SyRI è illegale, in quanto interferisce in modo sproporzionato sulla vita privata dei cittadini. Anche l'algoritmo utilizzato dal Governo italiano per assumere gli insegnanti in seguito alla riforma conosciuta come "Buona scuola" è stato dichiarato anticostituzionale: la decisione dell'algoritmo di assegnazione delle cattedre, infatti, non ha tutelato i diritti dei precari che venivano mandati a insegnare presso scuole lontane nonostante fossero disponibili posti di lavoro più vicini a casa loro. Come si legge nel rapporto *Getting the future right*: *artificial intelligence and fundamental rights* della FRA[3], la digitalizzazione dei sistemi di welfare è spesso accompagnata da riduzioni di budget, numero di beneficiari o altre misure che riducono il benessere sociale. La digitalizzazione, inoltre, offrendo la possibilità di controllare le persone, aumenta il potere degli stati e questo è particolarmente grave quando accade in paesi con deficit nello Stato di diritto. Sempre secondo il rapporto, l'utilizzo di algoritmi per gestire il welfare da parte della pubblica amministrazione desta non poche preoccupazioni rispetto all'impatto negativo che può avere sulle povertà e le disuguaglianze (pensiamo ai servizi di assistenza per l'infanzia o ai sussidi di disoccupazione). Attualmente,

[3] European Union Agency for Fundamental Rights, *Getting the future right: artificial intelligence and fundamental rights: report*, Publications Office, 2020, https://data.europa.eu/doi/10.2811/58563.

a livello globale, le nuove tecnologie vengono utilizzate in molti modi per amministrare sistemi di welfare (verifica dell'identità, valutazioni di ammissione, calcoli di benefici, prevenzione e rilevamento delle frodi ecc.) ed è quindi necessaria una riflessione sull'eticità e la legalità di queste tecnologie.

L'Intelligenza Artificiale avrà un notevole impatto anche sul mondo del lavoro. Secondo una stima del Parlamento Europeo, il 14% dei posti di lavoro nei paesi OCSE sono automatizzabili e il 32% potrebbe subire cambiamenti sostanziali nei prossimi anni[4]. Esistono anche problemi di concorrenza: se pensiamo che il dato ha acquisito valore economico è facile pensare che le società che possiedono più dati possono ottenere un grande vantaggio competitivo. In termini di sicurezza, invece, le applicazioni di IA che sono ad esempio a stretto contatto fisico con gli esseri umani possono comportare dei rischi se progettate male o utilizzate in modo inappropriato. Criticità emergono anche quando si considera la responsabilità legale: nel caso di incidenti con un'auto a guida autonoma o errori nell'ambito sanitario sarà difficile determinare se la responsabilità sia da attribuire al programmatore, al guidatore o al medico, oppure alla macchina in sé, soprattutto nel caso in cui il programmatore riesca a dimostrare che non sono stati fatti errori nella programmazione.

Viste le sfide poste dai sistemi IA, emerge il bisogno di un nuovo tipo di educazione e sensibilizzazione

[4] European Parliament News, *Artificial intelligence: threats and opportunities*, march 2021.

per i rischi associati a questi sistemi. Come vedremo in seguito, il **quadro normativo** attuale non è sufficiente per mitigare tutte le sfide etiche, legali e sociali sollevate dall'uso ormai pervasivo dell'Intelligenza Artificiale. Per porre rimedio all'ampio divario tra la velocità di avanzamento di queste tecnologie e la lentezza nello sviluppo normativo, negli ultimi anni alcune organizzazioni nazionali ed internazionali hanno sollevato la questione sull'eticità dell'IA, provando a definire alcuni principi etici. Anche l'Unione Europea si è posta l'obiettivo di promuovere un'IA antropocentrica delineando i suoi **principi etici** e pubblicando, ad aprile 2021, la sua prima **bozza di regolamentazione dell'IA**, nella quale viene proposto un approccio basato sul livello di rischio: i sistemi considerati con un livello di rischio alto dovranno sottostare ad una serie di obblighi, come l'essere sottoposti a **sistemi di valutazione e mitigazione del rischio**, prima di essere immessi sul mercato.

1.1. QUADRO GIURIDICO DI RIFERIMENTO

L'Intelligenza Artificiale non è una nuova tecnologia e, sebbene attualmente non esista un regolamento interamente dedicato ad essa, sono già in vigore alcuni provvedimenti normativi (il Regolamento generale sulla protezione dei dati e il Regolamento UE 2016/679 sulla libera circolazione dei dati non personali), la Convenzione 108/81, alcune direttive (ad esempio la direttiva macchine, la direttiva sulla responsabilità dei prodotti, il diritto dei consumatori, le direttive in materia di sicurezza e salute sul lavoro o norme specifiche

come il regolamento sui dispositivi medici nel settore sanitario) e numerose leggi degli Stati membri dell'UE che la disciplinano almeno in parte. Anche nell'ambito del diritto primario dell'UE esiste un corpus normativo giuridicamente vincolante per lo sviluppo, la distribuzione e l'utilizzo di IA: si tratta dei trattati dell'Unione Europea e della sua Carta dei diritti fondamentali. A tal proposito sono giuridicamente vincolanti anche i trattati ONU sui diritti umani e la Convenzione europea dei diritti dell'uomo. Nei prossimi paragrafi verranno velocemente trattati i **diritti fondamentali** (che ispirano in parte le linee guida per la progettazione di sistemi di IA affidabili) e, considerando che l'IA si nutre di molti **dati** personali, il Regolamento generale sulla protezione dei dati (**RGPD**) e la **Convenzione di Strasburgo** (Convenzione 108/81).

1.2. I DIRITTI FONDAMENTALI

Nella raccolta dei diritti fondamentali previsti dal diritto internazionale in materia di diritti umani, dai trattati UE e dalla Carta dei diritti fondamentali dell'UE, le seguenti famiglie di diritti sono particolarmente pertinenti per quanto riguarda i sistemi di IA (anche se non sempre contemplano una tutela giuridicamente completa):

- **Rispetto della dignità umana**: ogni essere umano possiede un valore intrinseco che non deve mai essere svalutato, compromesso o represso dagli altri e nemmeno dalle nuove tecnologie come i sistemi di IA. Essi devono quindi essere sviluppati in modo tale da rispettare, servire e proteggere l'integrità fi-

sica e psichica degli esseri umani, il senso di identità personale e culturale e la soddisfazione dei bisogni essenziali.

- **Libertà individuale**: ogni essere umano deve poter essere libero di prendere decisioni importati per sé stesso; pertanto, nell'ambito dell'IA, occorre ridurre al minimo la coercizione illegittima o indiretta, le minacce all'autonomia mentale e alla salute psichica, la sorveglianza ingiustificata, l'inganno e la manipolazione iniqua.
- **Rispetto della democrazia, della giustizia e dello Stato di diritto**: i sistemi di IA devono servire a mantenere e a promuovere i processi democratici e a rispettare i valori e le scelte di vita degli individui. Questi sistemi non devono compromettere i processi democratici, la decisione umana o i sistemi di voto democratico.
- **Uguaglianza, non discriminazione e solidarietà**: va garantito il rispetto per il valore morale e la dignità di tutti gli esseri umani. L'uguaglianza implica quindi che il funzionamento di un sistema di IA non possa portare a risultati ingiustamente distorti. Vanno inoltre tutelati maggiormente i gruppi e le persone potenzialmente vulnerabili.
- **Diritti dei cittadini**: i cittadini godono di molti diritti (di voto, di buona amministrazione, di accesso ai documenti pubblici, di presentare petizioni all'amministrazione ecc.). I sistemi di IA potrebbero avere effetti negativi su questi diritti che invece dovrebbero essere salvaguardati.

1.3. La protezione dei dati personali ai tempi dell'IA

Già a partire dagli anni '60, con il crescere delle tecnologie dell'informazione e della comunicazione, nasce l'esigenza di tutelare le persone rispetto al trattamento automatizzato dei dati. Proprio in questo contesto, nel 1981, il Consiglio d'Europa emana uno dei più importanti strumenti legali per la protezione delle persone rispetto all'automatizzazione del trattamento dei dati personali: la Convenzione di Strasburgo (108/1981). L'elaborazione automatizzata dei dati, infatti, è una grande risorsa di informazione: quando affermiamo che siamo maschi o femmine, o che viviamo a Monza invece che a Milano, non diamo soltanto dati relativi al sesso o al luogo, perché queste informazioni possono essere usate statisticamente per acquisire informazioni per arrivare alla profilazione. Il dato, infatti, apre sia un valore economico sia un valore politico: pensiamo al caso di Cambridge Analytica, dove i dati che pensavamo dicessero poco di noi o si perdessero nell'etere in realtà erano diventati dati di massa e, contenendo molte informazioni sulla società e sulle singole persone, venivano venduti e utilizzati per scopi politici (perché consentivano di attuare una campagna elettorale mirata). È chiaro che non vogliamo rinunciare all'elaborazione dei dati, perché è nel nostro interesse interagire e lasciare traccia delle nostre azioni (come ad esempio il comprare il biglietto di un concerto, di un treno o iscriverci a un evento). Con l'avanzare delle tecnologie e dell'immensa quantità di *big data* di cui le stesse hanno bisogno, il punto centrale diventa da un lato quello

di trovare un equilibrio tra circolazione e protezione dei dati personali, dall'altro quello di trovare strategie per ridurre al minimo i rischi associati a determinati utilizzi dell'Intelligenza Artificiale, senza però limitarne lo sviluppo tecnologico.

Il 24 maggio 2016 entra in vigore un importante regolamento relativo alla protezione dei dati personali e alla libera circolazione degli stessi (abrogando la Direttiva 95/46/CE): il Regolamento generale sulla protezione dei dati (Regolamento UE 2016/679). Il regolamento viene attuato nel 2018 e proprio nello stesso anno anche la Convenzione di Strasburgo viene aggiornata per avvicinarla al regime stabilito dallo stesso Regolamento generale sulla protezione dei dati (RGPD) e per rafforzare la protezione dei dati personali su larga scala.

È importante sottolineare che sebbene l'Intelligenza Artificiale non sia esplicitamente menzionata in questo quadro giuridico, molte disposizioni e principi sia del RGPD che della Convenzione sono rilevanti anche per i sistemi di Intelligenza Artificiale: questi sistemi, infatti, implicano il trattamento massiccio di dati personali e per questo non sono esonerati dall'adesione a questi principi. Analizzando nello specifico queste normative in relazione all'IA ci rendiamo però presto conto che forniscono buone basi per porre rimedio all'abuso dei dati personali, ma non risultano sufficienti a tutelare i cittadini di fronte all'ormai pervasivo utilizzo di questa tecnologia nella nostra società.

In particolare, alcuni **principi del RGPD** (Articolo 5), che ritroviamo anche all'Articolo 5 della Convenzione 108/81, si scontrano con la natura stessa di

un sistema di IA: pensiamo ad esempio all'attrito tra il **principio della limitazione delle finalità** (secondo cui i dati personali devono essere "raccolti per finalità determinate, esplicite e legittime, e successivamente trattati in modo che non sia incompatibile con tali finalità") e la possibilità che un sistema di IA riutilizzi quei dati per scopi diversi rispetto a quelli annunciati al momento della raccolta (anche in modo inconsapevole da parte del titolare del trattamento). Ugualmente è difficile applicare il **principio di minimizzazione del dato** (secondo cui i dati raccolti devono essere "adeguati, pertinenti e limitati a quanto necessario rispetto alle finalità per le quali sono trattati"): questi sistemi richiedono ormai un'enorme quantità di dati per poter apprendere al meglio e quindi non possono essere molto limitati. Anche il **principio di liceità, correttezza e trasparenza** (secondo cui i dati devono essere "trattati in modo lecito, corretto e trasparente nei confronti dell'interessato") risulta essere critico quando si tratta di sistemi basati sull'IA: l'Articolo 13, paragrafo 2, lettera f) richiede "trasparenza sull'esistenza di un processo decisionale automatizzato e sulla logica dello stesso", ma questi sistemi non sono basati su una logica umanamente comprensibile e quindi spesso anche gli stessi programmatori non riescono a spiegare la logica sottostante una decisione algoritmica (il cosiddetto fenomeno delle *black box*, ovvero dell'incapacità di spiegare il processo che ha portato a un *output* specifico). Risultano, inoltre, esserci criticità anche nel determinare la **responsabilità legale** nel caso di violazioni nel trattamento automatizzato di dati personali.

Un altro aspetto rilevante del RGPD in relazione ai sistemi di IA riguarda il **consenso** (Articolo 6): secondo il Regolamento, l'interessato al trattamento ha diritto ad essere informato sia sulle modalità che sulle finalità del trattamento dati, affinché il consenso fornito sia informato; questo meccanismo però non sembra essere sufficiente nella società digitale, proprio perché gli individui possono fornire il consenso senza comprendere appieno i termini e le condizioni specifiche di ogni processo di elaborazione dei dati, soprattutto quando esso riguarda l'utilizzo di algoritmi. Il consenso risulta essere debole non solo perché in realtà è "disinformato", ma anche perché spesso è condizionato e non esprimerlo significherebbe non poter accedere a molti servizi o opportunità. Se pensiamo, inoltre, che all'Articolo 22 del RGDP viene sostenuto il diritto a non essere sottoposti a decisioni esclusivamente algoritmiche e al paragrafo 2 lettera c) leggiamo però che questo diritto non si applica nel caso in cui la decisione "si basi sul consenso esplicito dell'interessato", possiamo comprendere appieno la debolezza del meccanismo del consenso.

Un altro limite del RGDP quando si tratta di sistemi di IA lo ritroviamo all'Art. 35, dove vengono richieste valutazioni dell'impatto esclusivamente per l'elaborazione dei dati personali (**DPIA**) e, in particolare solo per quei casi che comporteranno un rischio elevato per i diritti e le libertà delle persone fisiche. È quindi evidente che non verranno considerati molti casi di uso dell'IA ad alto rischio ma che non sono legati principalmente alla protezione dei dati.

Anche se in Europa l'IA non è così sviluppata come negli Stati Uniti o in Cina, il suo utilizzo sta aumentando in misura esponenziale ed è destinato a crescere sempre più, per questo motivo l'elaborazione normativa in tema di IA è necessaria. È infatti evidente l'ampio divario tra la velocità dello sviluppo tecnologico rispetto all'adeguamento normativo. Negli ultimi anni la Commissione Europea ha pubblicato alcuni documenti soprattutto di *soft law* o di autoregolazione, con l'obiettivo di incentivare un'IA antropocentrica; ad aprile 2021 la stessa Commissione ha pubblicato la sua prima proposta di Regolamentazione dell'IA con il duplice obiettivo di impedire che l'utilizzo di IA violi i diritti fondamentali e i valori europei, e assicurarsi che la regolamentazione non limiti lo sviluppo tecnologico, anzi lo promuova sviluppando un ecosistema di eccellenza e fiducia in questa tecnologia.

2. Principi etici e requisiti chiave per un'IA affidabile

In risposta alla crescente consapevolezza delle sfide indotte dall'Intelligenza Artificiale e alle relative lacune normative, nel corso degli ultimi anni è aumentato l'interesse verso l'eticità dell'Intelligenza Artificiale, definita anche come **IA responsabile**, **IA affidabile** o **IA benefica**. Indipendentemente dalla terminologia esatta, tutti questi inviti si riferiscono essenzialmente allo stesso obiettivo: il progresso dell'IA tale che i suoi benefici siano massimizzati mentre i suoi rischi e pericoli siano mitigati o prevenuti. Organizzazioni internazionali e istituzioni in questi ultimi anni hanno nominato comitati di esperti ad hoc sull'IA con lo scopo di redigere rapporti e documenti di orientamento sull'IA. In particolare, la Commissione Europea, i singoli stati, organizzazioni quali l'Organizzazione per la cooperazione e lo sviluppo economico (OCSE), l'Organizzazione delle Nazioni Unite per l'educazione, la scienza e la cultura (UNESCO), e istituti come l'*Institute of Electrical and Electronics Engineer* (IEEE) hanno delineato **linee guida e principi etici** non vincolanti sull'IA. Sforzi simili sono in questo momento compiuti anche aziende private quali BMW, Vodafone, Microsoft, Google, Accenture, IBM, Sony.

Rimanendo nel contesto europeo, nel 2019 l'*High-Level Expert Group* (HLEG), ovvero il Gruppo di esperti di alto livello sull'Intelligenza Artificiale nominato dalla Commissione Europea, ha pubblicato il do-

cumento "**Orientamenti etici per un'IA affidabile**"[1]. Secondo la Commissione, infatti, gli esseri umani e le comunità avranno fiducia nello sviluppo e nell'applicazione delle tecnologie soltanto quando esisterà un quadro di riferimento chiaro e completo. Proprio perché i sistemi di IA possono presentare dei rischi è necessario prevenirli massimizzandone i benefici. A tal fine è necessario realizzare sistemi di **IA antropocentrici**, ovvero al servizio dell'umanità e del bene comune, con l'obiettivo di migliorare il benessere e la libertà degli individui. Secondo la Commissione, per le persone e la società l'affidabilità dell'IA rappresenta un prerequisito per lo sviluppo, la distribuzione e l'utilizzo di sistemi della stessa, perché "se i sistemi di IA – e gli esseri umani che li creano – non sono degni di fiducia senza alcun dubbio, possono verificarsi conseguenze indesiderate, e di conseguenza l'adozione dell'IA potrebbe essere ostacolata, impedendo la realizzazione dei benefici sociali ed economici potenzialmente enormi apportati da tali sistemi". Un sistema di IA, per essere definito affidabile, durante l'intero ciclo di vita deve possedere le seguenti componenti:

- **Legalità**: rispetto di tutte le leggi e di tutti i regolamenti applicabili.
- **Eticità**: adesione a principi e valori etici.
- **Robustezza**: sia dal punto di vista tecnico sia dal punto di vista sociale, perché i sistemi di IA possono causare danni non intenzionali.

[1] Commissione Europea, *Orientamenti etici per un'IA affidabile*, 2019.

Negli Orientamenti non viene trattata la prima componente (legalità), bensì vengono offerte indicazioni per promuovere e garantire l'eticità e la robustezza dei sistemi di IA. L'approccio all'etica dell'IA è basato sui diritti fondamentali stabiliti dai trattati UE, dalla Carta dei diritti fondamentali dell'UE, e dal diritto internazionale in materia di diritti umani. Negli Orientamenti, infatti, vengono definiti i **quattro principi etici** per un'IA affidabile proprio a partire dai cinque diritti fondamentali (descritti nel paragrafo 2.1) ai quali bisogna aderire per garantire che i sistemi di IA siano sviluppati, distribuiti e utilizzati in modo affidabile:

- **Principio del rispetto dell'autonomia umana**: vanno garantiti il rispetto della libertà e dell'autonomia degli esseri umani. I sistemi di IA non devono "subordinare, costringere, ingannare, manipolare, condizionare o aggregare in modo ingiustificato gli esseri umani", ma piuttosto devono "essere progettati per aumentare, integrare e potenziare le abilità cognitive, sociali e culturali umane". Inoltre, secondo questo principio, la distribuzione delle funzioni tra esseri umani e sistemi di IA dovrebbe seguire i principi di progettazione antropocentrica, lasciando ampie opportunità di scelta all'essere umano (garantendo la sorveglianza e il controllo di tutti i processi operativi dei sistemi di IA).
- **Principio di prevenzione dei danni**: i sistemi di IA non devono arrecare nessun danno fisico o psichico, o influenzare in alcun modo negativamente gli esseri umani; per questo motivo devono essere tecnicamente solidi, garantendo così che non sia-

no esposti a usi malevoli. Inoltre, occorre prestare attenzione 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". Il principio di prevenzione dei danni presuppone anche il rispetto dell'ambiente naturale e di tutti gli esseri viventi.

- **Principio di equità**: secondo questo principio, lo sviluppo, la distribuzione e l'utilizzo dei sistemi di IA devono essere equi. Il Gruppo di esperti di alto livello dell'IA (HLEG AI) considera sia la dimensione sostanziale sia la dimensione procedurale dell'equità. Per la prima è necessario garantire una distribuzione giusta ed equa di costi e benefici, garantire libertà da distorsioni inique, discriminazioni e stigmatizzazioni, promuovere le pari opportunità in termini di accesso all'istruzione, ai beni, ai servizi e alla tecnologia, e fare in modo che i sistemi di IA non ingannino gli utenti né ne ostacolino la libertà di scelta. Per la seconda s'intende la capacità di impugnare le decisioni elaborate dai sistemi di IA e la possibilità di presentare un ricorso efficace contro di esse.
- **Principio di esplicabilità**: questo principio è fondamentale per mantenere la fiducia degli utenti nei confronti dei sistemi di IA. I processi devono essere trasparenti, le decisioni devono poter essere spiegate a coloro che ne sono interessati, le capacità e lo scopo dei sistemi di IA devono essere comunicati apertamente. Il Gruppo di esperti di alto livello sot-

tolinea comunque che non sempre è possibile poter spiegare perché un modello ha generato una particolare decisione (effetto scatola nera) e suggerisce altre misure per garantire l'esplicabilità (tracciabilità, verificabilità e la comunicazione trasparente sulle capacità del sistema).

I quattro principi vengono poi tradotti nei seguenti sette requisiti chiave (approfonditi nei prossimi paragrafi), che un sistema di IA dovrebbe soddisfare per essere considerato affidabile:

- Intervento e sorveglianza umani.
- Robustezza tecnica e sicurezza.
- Riservatezza e *governance* dei dati.
- Trasparenza.
- Diversità, non discriminazione ed equità.
- Benessere sociale e ambientale.
- *Accountability*.

2.1. Requisito di intervento e sorveglianza umani

Secondo il requisito di intervento e sorveglianza umani, che rientra nel principio di rispetto dell'autonomia umana, i sistemi di IA dovrebbero sostenere l'autonomia e il processo decisionale degli esseri umani. Di conseguenza i sistemi di IA dovrebbero:

Promuovere i **diritti fondamentali:** questi sistemi possono sia agevolare sia influire negativamente sui diritti fondamentali; se esistono rischi di questo tipo è consigliato eseguire una valutazione dell'impatto sui

diritti fondamentali prima dello sviluppo del sistema, per individuare e ridurre al minimo i rischi al fine di rispettare diritti e libertà individuali.

Sostenere l'**intervento umano:** questi sistemi possono essere utilizzati per influenzare il comportamento umano (attraverso manipolazioni inique, inganni, frodi ecc.) di conseguenza è importante che agli utenti vengano fornite conoscenze e strumenti per comprendere e interagire con i sistemi IA. In questo modo saranno in grado di prendere decisioni informate, in piena autonomia e di contestare il sistema se necessario.

Consentire la **sorveglianza umana:** per evitare che un sistema di IA non provochi effetti negativi o comprometta l'autonomia umana è necessario garantire l'intervento, la supervisione o il controllo da parte degli umani; si può anche prendere la decisione di non utilizzare sistemi di IA in situazioni critiche, o prevedere la possibilità di ignorare la decisione presa dal sistema.

2.2. REQUISITO DI ROBUSTEZZA TECNICA E SICUREZZA

Secondo questo requisito, che rientra nel principio di prevenzione dei danni, una componente fondamentale per assicurarsi che l'IA sia affidabile è la robustezza tecnica; pertanto si dovrebbero garantire:

- La **resilienza agli attacchi informatici**: i sistemi di IA devono essere protetti contro le vulnerabilità che possono esporli ad attacchi informatici (che possono colpire i dati, il modello o l'infrastruttura sottostante sia hardware che software).
- Un **piano di emergenza e sicurezza generale**: i si-

stemi di IA devono essere provvisti di funzionalità per attivare piani di emergenza in caso di problemi (ad esempio richiedendo l'intervento dell'operatore umano oppure passando da una procedura statistica a una basata su regole). È necessario ridurre al minimo le conseguenze di errori non intenzionali e valutare potenziali rischi associati all'uso di questi sistemi testando eventuali misure di sicurezza proattivamente.

- La **precisione**: soprattutto quando le decisioni prese dal sistema di IA influiscono direttamente sulla vita delle persone, è necessario garantirne la precisione (capacità di formulare un giudizio corretto, fare previsioni precise, adottare decisioni esatte sulla base dei dati o dei modelli).
- L'**affidabilità** e la **riproducibilità**: un sistema di IA è considerato affidabile quando funziona correttamente con una serie di *input* in diverse situazioni, ed è considerato riproducibile quando mostra lo stesso comportamento se ripetuto nelle medesime condizioni (questo consente di descrivere accuratamente il comportamento del sistema di IA).

2.3. Requisito di riservatezza e governance dei dati

Questo requisito, che rientra nel principio di prevenzione dei danni, riguarda il diritto fondamentale della riservatezza (diritto tutelato anche dal RGDP) e quindi è auspicabile:

- Garantire la **riservatezza** e la **protezione dei dati** durante l'intero ciclo di vita del sistema: sia le infor-

mazioni fornite dall'utente sia quelle generate nel corso dell'interazione con il sistema devono essere protette e non devono essere utilizzate ai fini di un'illecita o iniqua discriminazione;
- Garantire la **qualità** e l'**integrità dei dati**: prima di addestrare la macchina con un set di dati è necessario controllare che sia di qualità e non contenga distorsioni, imprecisioni o errori. I processi e i set di dati devono essere testati e documentati in ogni fase ed è inoltre necessario garantire l'integrità di questi dati;
- Regolare l'**accesso ai dati**: è necessario indicare, tramite adeguati protocolli, chi può accedere ai dati e in quali circostanze.

2.4. REQUISITO DI TRASPARENZA

Secondo questo requisito, che è connesso al principio dell'esplicabilità, dati, sistema e modelli di business devono essere trasparenti; di conseguenza è necessario garantire:
- La **tracciabilità**: è necessario documentare secondo i migliori standard sia i dati (compresi quelli di etichettatura e raccolta) sia gli algoritmi utilizzati e i processi che hanno portato il sistema di IA ad una decisione. Questo processo di documentazione facilita sia la spiegabilità che la verificabilità;
- La **spiegabilità**: affinché possa essere garantita è necessario essere in grado di spiegare sia i processi tecnici di un sistema di IA sia le relative decisioni umane, in modo tale che gli esseri umani possano capire e tener traccia delle decisioni prese dal si-

stema. Non è sempre facile poter spiegare come il sistema sia arrivato ad una decisione; pertanto, sarà necessario trovare un compromesso tra la spiegabilità (a discapito della precisione) e l'aumento di precisione (a discapito della spiegabilità). Tuttavia, nel caso in cui il sistema influisca sulla vita delle persone, deve essere sempre possibile fornire spiegazioni adeguate sul suo processo decisionale;
- La **comunicazione** agli esseri umani sul fatto che stanno interagendo con un sistema di IA.

2.5. Requisito di diversità, non discriminazione ed equità

Secondo questo requisito, che rientra nel principio di equità, è necessario che durante l'intero ciclo di vita del sistema siano permesse l'inclusività e la diversità; di conseguenza è necessario:
- **Evitare distorsioni inique**: siccome i set di dati utilizzati dai sistemi di IA possono essere incompleti o influenzati da distorsioni storiche non intenzionali (portando determinati gruppi o persone ad essere oggetto di pregiudizi o discriminazioni) le distorsioni identificabili e discriminatorie dovrebbero essere eliminate già nella fase di raccolta dei dati. Andrebbero inoltre attuati processi di sorveglianza per analizzare le finalità, i vincoli, i requisiti e le decisioni del sistema al fine di evitare distorsioni inique durante lo sviluppo del sistema di IA;
- Garantire l'**accessibilità** e **progettazione universale**: i sistemi di IA dovrebbero essere incentrati

sull'utente e progettati in modo che tutte le persone possano usare prodotti e servizi di IA. È auspicabile, inoltre, garantire l'accessibilità alla tecnologia anche alle persone con disabilità;
- Garantire la **partecipazione degli stakeholders** durante l'intero ciclo di vita del sistema di IA.

2.6. REQUISITO DI BENESSERE SOCIALE E AMBIENTALE

Questo requisito è in linea sia con il principio di equità sia con quello di prevenzione dei danni. L'IA dovrebbe essere utilizzata a vantaggio di tutti gli esseri umani, comprese le generazioni future, pertanto:
- I sistemi di IA devono essere **sostenibili** e rispettosi dell'ambiente: il processo di sviluppo, distribuzione, utilizzo del sistema e l'intera catena di approvvigionamento dovrebbero essere valutati in modo che il sistema sia il più possibile rispettoso dell'ambiente;
- Dev'essere considerato l'**impatto** che i sistemi di IA possono avere **sulla società**: possono infatti influenzare le relazioni sociali e i legami affettivi e alterare la concezione di intervento sociale, migliorando le abilità sociali ma anche contribuendo al loro deterioramento (pensiamo alla dequalificazione sul lavoro).

2.7. REQUISITO DI ACCOUNTABILITY

Questo requisito, che rientra nel principio di equità, richiede che vengano attuati meccanismi che garantiscano l'accountability dei sistemi di IA e dei loro risultati, pertanto è necessario garantire:

- La **verificabilità**, ovvero la possibilità di valutare algoritmi, dati e processi di progettazione. Le valutazioni e le relative relazioni da parte di revisori tanto esterni quanto interni possono aiutare ad aumentare l'affidabilità dell'IA. Quando le decisioni dei sistemi di IA influiscono sui diritti fondamentali, inoltre, sarebbe auspicabile sottoporle a una verifica indipendente.
- La **riduzione degli effetti negativi**: si deve essere in grado di riferire le azioni o le decisioni che hanno contribuito ad un determinato risultato. Valutazioni d'impatto potrebbero essere utili per ridurre al minimo gli effetti negativi del sistema di IA;
- La **motivazione sul compromesso accettato**: come indicato nel requisito di trasparenza, talvolta possono insorgere tensioni tra i requisiti ed è inevitabile dover raggiungere un compromesso. Ciò implica la valutazione in termini di rischio per i princìpi etici e i diritti fondamentali (nel caso in cui non si riesca a trovare un compromesso eticamente accettabile è consigliabile abbandonare il modello) e l'analisi della documentazione relativa alle motivazioni della decisione presa;
- Il **ricorso** attraverso meccanismi accessibili e adeguati in caso di effetti negativi ingiusti.

2.8. Limiti dei principi e requisiti etici

I principi e requisiti sopraelencati offrono delle linee guida da seguire affinché un sistema di IA sia considerato affidabile, tuttavia mostrano alcune limitazioni. Come affermato negli stessi Orientamenti etici per

un'IA affidabile, possono ad esempio esserci **tensioni tra i principi**: pensiamo ad esempio alla spiegabilità del sistema di IA che può entrare in conflitto con la precisione del sistema stesso (un sistema più preciso può essere più difficile da spiegare). La necessità di set di dati sempre più grandi e diversificati sembra difficile da conciliare con i requisiti di riservatezza e di autonomia umana. Inoltre, i principi più frequentemente citati sono anche quelli per i quali possono essere o sono già stati sviluppati accorgimenti tecnici: accountability, spiegabilità, privacy, equità o robustezza tecnica e sicurezza sono più facilmente operativi matematicamente e quindi tendono ad essere implementati in termini di soluzioni tecniche.

Gli orientamenti etici, inoltre, non hanno meccanismi di applicazione che vadano oltre **l'adesione volontaria** e numerose aziende sono desiderose di utilizzare l'IA per scopi redditizi in molteplici modi: questi tipi di utilizzo non sono inquadrati in un'etica basata su valori e principi, ma piuttosto collocati all'interno di una logica principalmente economica. Ingegneri e sviluppatori spesso non sono istruiti sulle questioni etiche e non hanno comunque il potere di sollevarle: nei contesti aziendali saltare le considerazioni etiche equivale a prendere il percorso più veloce, di conseguenza sviluppo, implementazione e uso dei sistemi di IA hanno spesso poco a che fare con i principi etici.

Proprio perché le linee guida e i principi etici non sono sufficienti, analogamente a quanto richiesto nel RGDP per la protezione dei dati, organizzazioni internazionali suggeriscono un **approccio basato sul rischio**

e valutazioni dell'impatto anche per i sistemi di IA, in particolare: l'UNESCO incoraggia gli Stati membri ad introdurre valutazioni dell'impatto dei sistemi di IA per valutare vantaggi e preoccupazioni, individuando e prevenendo i rischi[2]; e l'OECD suggerisce una gestione del rischio in ciascuna fase del ciclo di vita di un sistema di IA (Art 1.4)[3]. Anche l'IEEE in "Ethically aligned design"[4] sottolinea l'importanza di un'analisi del rischio per i sistemi autonomi intelligenti e suggerisce una valutazione dell'impatto. Come vedremo nel prossimo capitolo anche la Commissione Europea, nella sua proposta di regolamentazione dell'IA, propone un approccio basato sul rischio e suggerisce attente valutazioni del rischio e mitigazioni dell'impatto da effettuare prima di distribuire sistemi di IA definiti ad alto rischio.

[2] UNESCO, *Preliminary report on the first draft of the Recommendation on the Ethics of Artificial Intelligence*, 2020.
[3] OECD, *Recommendation of the council on artificial intelligence*, 2019.
[4] IEEE, *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*, 2019.

3. La proposta di regolamentazione dell'IA della Commissione Europea

Come precedentemente accennato, nell'aprile 2021 la Commissione Europea ha proposto il suo primo quadro giuridico sull'Intelligenza Artificiale[1] con lo scopo di stabilire un quadro giuridico uniforme per il miglioramento del funzionamento del mercato interno, in particolare per lo sviluppo, il marketing e l'uso dell'IA in conformità con i valori dell'Unione. Alcuni stati membri hanno già esaminato l'adozione di norme nazionali per garantire che l'Intelligenza Artificiale sia sviluppata e utilizzata in conformità con obblighi in materia dei diritti fondamentali. Proprio perché norme nazionali diverse possono portare alla frammentazione del mercato interno e ad incertezze giuridiche per sviluppatori o utilizzatori di sistemi di IA, si legge nella proposta, è necessario stabilire obblighi uniformi per tutelare i diritti delle persone in tutto il mondo.

Analogamente a quanto suggerito dall'IEEE, l'OECD e l'UNESCO, la proposta segue un **approccio basato** sul rischio e differenzia gli usi di IA sulla base di quei rischi definiti inaccettabili, elevati o bassi. Sono classificati, ad esempio, a **rischio inaccettabile** e quindi saranno vietati: i sistemi di punteggio sociale per determinare l'affidabilità di persone e imprese; i sistemi

[1] European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence act) and amending certain union legislative acts*, april 2021.

utilizzati per la sorveglianza di massa; e i sistemi che manipolano il comportamento umano con tecniche subliminali o sfruttano le vulnerabilità di specifichi gruppi causando danni fisici o psicologici.

Sono, invece, considerati ad **alto rischio** quei sistemi di IA che possono presentare un rischio di danno alla salute o alla sicurezza o un rischio di impatto negativo sui diritti fondamentali. Nello specifico, all'Allegato III della proposta sono elencati una serie di sistemi di IA considerati ad alto rischio, quali: sistemi destinati ad essere utilizzati per la biometria remota in tempo reale; sistemi preposti alla sicurezza nella gestione della circolazione stradale e di fornitura di acqua, gas, riscaldamento ed elettricità; sistemi che determinano l'accesso o assegnazione di persone fisiche a istituti di istruzione e formazione professionale; sistemi destinati ad essere utilizzati allo scopo di valutare gli studenti in istituti di istruzione e formazione professionale oppure sistemi utilizzati per valutare i partecipanti a test di ammissione scolastica; sistemi utilizzati per il reclutamento o la selezione di persone fisiche, oppure per la promozione o la cessazione di rapporti contrattuali di lavoro; sistemi destinati ad essere utilizzati dalle autorità pubbliche per valutare l'idoneità delle persone fisiche all'assistenza pubblica, a benefici e servizi, nonché quelli utilizzati per concedere, ridurre revocare o reclamare vantaggi e servizi; sistemi impiegati per stabilire la priorità o inviare la richiesta di primo intervento durante un'emergenza; sistemi utilizzati dalle autorità per valutare sia il rischio di commissione di reati o recidive sia il rischio per potenziali vittime di reati; sistemi

utilizzati dalle autorità per rilevare lo stato emotivo di una persona fisica o per valutare l'attendibilità delle prove nel corso delle indagini e dei procedimenti giudiziari; sistemi adoperati dalle autorità per valutare un rischio per la sicurezza o per la salute di una persona che entra o entrerà nel territorio di uno Stato membro o per valutare un rischio di immigrazione irregolare; sistemi di IA destinati ad essere utilizzati dalle autorità per verificare l'autenticità dei documenti di viaggio e supporto o per esaminare domande d'asilo, visti, permessi di soggiorno.

I sistemi di IA che rientrano nella categoria dei sistemi ad alto rischio saranno soggetti a severi obblighi, ai quali dovranno conformarsi prima di poter essere immessi sul mercato. Dovranno garantire un'elevata qualità dei set di dati, registrare le attività per garantire la tracciabilità dei risultati, documentare dettagliatamente le informazioni necessarie sul sistema e il suo scopo affinché le autorità possano valutarne la conformità; fornire informazioni chiare e adeguate agli utenti. Inoltre, deve essere presente un adeguato controllo umano e un elevato livello di robustezza, sicurezza e precisione. Infine, dovranno adottare adeguati sistemi di valutazione e mitigazione del rischio.

I sistemi di Intelligenza Artificiale classificati a **basso rischio**, invece, dovranno rispettare gli obblighi di trasparenza (ad esempio gli utenti devono essere consapevoli che stanno interagendo con una macchina e non con una persona reale in modo tale da poter prendere una decisione informata). La proposta non disciplina quei sistemi di IA che presentano **rischi minimi**

o nulli per i cittadini europei.

Presto quindi potranno essere banditi alcuni utilizzi dei sistemi di IA definiti a rischio inaccettabile e molte tecnologie già in uso in Europa (ad esempio quelle utilizzate nella pubblica amministrazione o per scansionare i curriculum) saranno definite ad alto rischio e dovranno soddisfare determinati standard per continuare ad essere utilizzate.

4. Sistemi di valutazione dell'impatto dell'IA

Come brevemente accennato nel capitolo 3, esistono molti esempi di linee guida non vincolanti analoghi a quelli delineati dalla Commissione Europea. Questi documenti generalmente non contengono linee guida chiare sulla valutazione dell'impatto, bensì evidenziano i diversi criteri e principi che dovrebbero essere presi in considerazione nell'elaborazione e nell'esecuzione di una valutazione dell'impatto del sistema di IA. Tale valutazione funziona in modo simile alla DPIA richiesta all'Articolo 35 del RGDP e in particolare è utile per aiutare sviluppatori, distributori e coloro che desiderano utilizzare sistemi di Intelligenza Artificiale ad essere consapevoli dei rischi che possono comportare questi sistemi. Consente, inoltre, di **valutare l'impatto** del sistema sugli aspetti etici e legali con lo scopo di **mitigare potenziali rischi**. Generalmente è consigliato effettuare la valutazione dell'impatto sia durante lo sviluppo del sistema sia prima della sua distribuzione (per prevenire in tempo errori che potrebbero essere in seguito costosi), ma anche periodicamente per valutare le conseguenze del sistema di Intelligenza Artificiale nel tempo.

Sono disponibili diversi strumenti pratici (liste di controllo, elenchi di domande, strumenti di autovalutazione online, strumenti di gestione del rischio) per valutare l'impatto delle tecnologie dell'IA e mitigarne i rischi. Nel luglio 2020, ad esempio, il Gruppo di esperti di alto livello sull'Intelligenza Artificiale ha pubblicato

"Assessment list for Trustworthy AI (ALTAI)"[1], ovvero un elenco di domande per aiutare le organizzazioni a valutare l'affidabilità (in base ai quattro principi e sette requisiti elencati negli Orientamenti etici della Commissione Europea) del loro sistema riducendo i potenziali rischi derivanti dall'uso di questa tecnologia. Altri esempi sono: l'Ethics Toolkit[2], ovvero uno strumento opensource basato su un approccio di gestione del rischio e progettato per supportare i governi locali; la valutazione proposta dal World Economic Forum[3] basata sul rischio e costituita da una *checklist* di domande per aiutare a valutare l'impatto dei sistemi di IA; l'articolata "Artificial Intelligence Impact Assessment" proposta dall'Electronic Commerce Platform Nederland (ECP)[4]; la valutazione dell'impatto proposta da AlgorithmWatch[5] basata su due liste di controllo che guidano nella redazione di una relazione sulla trasparenza (per dimostrare che sono state considerate le questioni etiche più rilevanti); e il framework propo-

---

[1] European Commission, *Assessment list for trustworthy AI (ALTAI)*, 2019.

[2] Center of Government Excellence, Johns Hopkins University, *Ethics & Algorithm toolkit*, 2018.

[3] World Economic Forum, *AI Procurement in a Box*, Workbook, 2020.

4ECP (Platform voor de InformatieSamenleving), *Artificial intelligence impact assessment*, 2018.

[5] Michele Loi, Anna Mätzener, Angela Müller and Matthias Spielkamp, *Automated Decision Making Systems in the Public Sector. An Impact Assessment Tool for Public Authorities*, Algorithm Watch, june 2021.

sto dall'istituto AI Now[6] che comprende la spiegazione del processo e degli aspetti da prendere in considerazione per effettuare una valutazione dell'impatto del sistema di IA.

Le valutazioni dell'impatto del sistema di IA non sono obbligatorie e non costituiscono alcun onere amministrativo, ma al contrario **sono un supporto per la diffusione responsabile di questa tecnologia**. Al di là dei requisiti richiesti dal RGPD, infatti, considerando che ad oggi quella della Commissione Europea rimane una proposta, esistono pochi esempi di leggi che richiedono valutazioni obbligatorie sugli effetti dell'Intelligenza Artificiale. Vista la crescente diffusione dell'IA, il Governo canadese, ad esempio, ha pubblicato le linee guida e i requisiti obbligatori per valutare l'impatto dell'IA nel settore della pubblica amministrazione: ai sensi della direttiva canadese "Directive on Automated Decision-Making (DADM)" è, infatti, obbligatorio effettuare la valutazione per qualsiasi sistema, strumento o modello statistico utilizzato per prendere una decisione amministrativa.

4.1. Esempi di valutazione dell'impatto

Di seguito vengono proposti tre esempi di strumenti di valutazione dell'impatto dei sistemi di IA: l'**Algorithmic Impact Assessment (AIA)** del Governo canadese[7], l'**Artificial Intelligence Impact Assessment**

[6] Dillon Reisman, Jason Schultz, Kate Crawford and Meredith Whittaker, *Algorithmic Impact Assessments: a practical framework for public agency accountability*, AI Now, 2018.
[7] Government of Canada, Algorithmic Impact Assessment

(AIIA) di ECP e l'**Impact Assessment Tool** di AlgorithmWatch pubblicato a giugno 2021.

Algorithmic Impact Assessment (AIA)

L'Alghorithmic Impact Assessment è uno strumento obbligatorio di valutazione del rischio ai sensi della "Directive on Automated Decision-Making (DADM)". La direttiva è entrata in vigore il 1 aprile 2019 con l'obiettivo di garantire "*that Automated Decision Systems are deployed in a manner that reduces risks to Canadians and federal institutions, and leads to more efficient, accurate, consistent, and interpretable decisions made pursuant to Canadian law*".

L'AIA è progettata per aiutare i dipartimenti e le agenzie a comprendere e gestire meglio i rischi associati ai sistemi decisionali automatizzati e rappresenta una valutazione automatizzata composta da quarantotto domande per valutare il rischio del sistema e da trentatré domande di mitigazione. Le domande prevedono risposte su scala dicotomica o nominale e sono progettate per misurare l'impatto che le decisioni avranno su diversi fattori, tra cui: i diritti degli individui o delle comunità, la salute o il benessere di individui o comunità, gli interessi economici di individui, entità o comunità, la continua sostenibilità di un ecosistema; la durata e la reversibilità degli impatti.

Per **definire il rischio**, ad esempio, vengono poste domande relative: al progetto; alla tipologia di sistema, algoritmo e decisione automatizzata; all'impatto

---

Tool, 2019.

su libertà, salute, economia e ambiente; al tipo di dati utilizzati e alla loro fonte (figura 9).

Per **valutare le misure di mitigazione** in atto vengono poste domande relative: agli stakeholders consultati (ad esempio esperti in materia di privacy); alla qualità dei dati (che siano rappresentativi e imparziali); alle procedure in atto per controllare il sistema e le sue decisioni; alle misure prese per salvaguardare le informazioni personali.

Ad ogni domanda, in base alla risposta data, viene attribuito un punteggio. Una volta terminato l'AIA vengono restituiti i risultati che comprendono il livello di impatto del sistema di IA (su quattro livelli) e un collegamento ai requisiti della Direttiva.

Artificial Intelligence Impact Assessment (AIIA)

Nel contesto europeo, l'Electronic Commerce Platform Nederland (ECP) propone il suo strumento di valutazione dell'impatto dell'Intelligenza Artificiale, ovvero un metodo strutturato per:

- mappare i vantaggi pubblici di un'applicazione di IA;
- analizzare l'affidabilità, la sicurezza e la trasparenza;
- identificare valori e interessi in gioco nella distribuzione di IA;
- identificare e limitare i rischi legati alla diffusione di IA;
- tenere conto delle scelte effettuate nell'identificazione di valori e interessi.

L'AIIA è composta da otto step. Il primo step consiste in una serie di domande di screening per determinare se sia utile o meno fare un'AIIA. Le domande

riguardano il contesto sociale e politico dell'applicazione, le caratteristiche della stessa (in termini di autonomia, complessità, comprensibilità, prevedibilità) e i processi di cui fa parte (in termini di complessità dell'ambiente e processo decisionale, trasparenza, comprensibilità, prevedibilità dei risultati, impatto per le persone). Se si risponde in mondo affermativo a una delle domande proposte nel primo step viene consigliato di sottoporre il sistema a una valutazione dell'impatto. La valutazione vera e propria inizia dalla fase due, dove viene richiesta la descrizione del progetto (informazioni sugli obiettivi, sulla tecnologia, sui dati utilizzati, sugli stakeholders). La fase tre è dedicata ai benefici che l'utilizzo di IA comporta. Nella fase quattro, invece, viene richiesto di descrivere l'influenza che può avere la tecnologia sui valori umani e sociali, considerando i seguenti principi: dignità umana, autonomia umana, responsabilità, trasparenza, equità, democrazia, sicurezza, privacy e protezione dei dati e sostenibilità. Per quanto riguarda gli altri step: lo step cinque è dedicato a valutare se l'applicazione di IA è affidabile, sicura e trasparente; lo step sei è dedicato a considerare congiuntamente i benefici della fase tre e i rischi della fase quattro; lo step sette è dedicato alla corretta registrazione degli step da due a cinque e dalla documentazione delle decisioni prese; l'ultimo step è dedicato al monitoraggio e alla valutazione dell'impatto del sistema.

L'AIIA di ECP propone una valutazione fornendo domande guida aperte: da un lato questa modalità permette di indagare con più precisione e a fondo il siste-

ma di IA, dall'altra il procedimento può essere lungo in termini di tempistiche (soprattutto se consideriamo il fatto che è consigliato eseguire la valutazione dell'impatto periodicamente). D'altra parte, le domande con risposta su scala dicotomica dell'AIA canadese non permettono di rilevare le diverse sfaccettature: considerando ad esempio la domanda "*The algorithmic process will be difficult to interpret or to explain*", la difficoltà di spiegazione di un algoritmo può essere misurata su diversi intervalli e non semplicemente come assenza o presenza di difficoltà. Interessante è, invece, la valutazione pubblicata a giugno di quest'anno da AlgorithmWatch.

Impact Assessment Tool for Public Authorities

In collaborazione con l'Università di Basilea, AlgorithmWatch ha condotto uno studio sull'uso dell'Intelligenza Artificiale nella pubblica amministrazione su richiesta del Canton Zurigo. A seguito di questo studio, è stato sviluppato uno strumento concreto e praticabile di valutazione dell'impatto di specifici sistemi decisionali automatizzati applicati dalle autorità pubbliche. Basandosi principalmente sui quattro principi e sette requisiti etici definiti dal "HLEG AI- High-level Expert Group on Artificial Intelligence" nominato dalla Commissione Europea, è stata introdotta una procedura di valutazione dell'impatto definita in due fasi. Nella prima fase viene proposta una **lista di controllo (triage)** dove l'amministrazione deve valutare quali questioni di trasparenza etica valga la pena documentare in dettaglio durante l'esecuzione del progetto.

Nel caso in cui il sistema preso in esame dovesse essere soggetto a ulteriori requisiti di trasparenza le autorità pubbliche dovranno **redigere una relazione completa** sulla trasparenza. Nella seconda fase, infatti, viene proposta una lista di controllo con lo scopo di fornire una guida per la stesura di una relazione sulla trasparenza altamente dettagliata.

AlgorithmWatch raccomanda la stesura del **rapporto della trasparenza** già dalle prime fasi di avvio del progetto (questo perché alcune delle informazioni richieste per il rapporto sulla trasparenza possono essere generate solo durante le diverse fasi dell'esecuzione del progetto e non possono più essere generate quando il sistema è ormai ultimato). Per questo motivo l'elenco di controllo della relazione sulla trasparenza include anche indicazioni sulle fasi del processo in cui devono essere generate le informazioni specifiche necessarie per la trasparenza. Terminato il processo, il rapporto sulla trasparenza dovrà includere informazioni chiare sui meccanismi messi in atto per affrontare le specifiche questioni etiche evidenziate nel primo elenco di controllo (triage). Inoltre, se i meccanismi verranno modificati dopo l'implementazione del sistema, l'amministrazione dovrà verificare se la valutazione iniziale è ancora valida o se sono emersi ulteriori problemi di trasparenza etica.

Nel caso in cui l'amministrazione non sia in grado di fornire un adeguato grado di trasparenza sulle questioni etiche, o se l'esito dei meccanismi di trasparenza dimostra l'inadeguatezza del sistema, è necessario rivalutare l'obiettivo del progetto e considerare l'inve-

stimento di maggiori risorse per trovare una soluzione consona a rendere il sistema etico. In questa proposta di valutazione dei sistemi di IA, la trasparenza è intesa come una comunicazione a diversi tipi di pubblico (ovvero coloro che hanno il diritto ad accedere alle informazioni). Sempre secondo AlgorithmWatch, per ottenere una trasparenza significativa, tutti i sistemi decisionali automatizzati impiegati nel settore pubblico dovrebbero essere divulgati all'interno di un registro pubblico. Questi registri dovrebbero inoltre contenere informazioni sullo scopo, sugli stakeholders e sui risultati della valutazione dell'impatto (in questo caso la relazione sulla trasparenza). La prima lista di controllo dovrebbe essere stilata già durante la fase di pianificazione del sistema, così da poter prendere in considerazione, oltre all'obiettivo, altre specifiche aggiuntive per il progetto. L'elenco di controllo consente di determinare i principali problemi di trasparenza che devono essere documentati nella relazione ed è composto dalle domande raccolte nella seguente tabella:

1.1.	Does the decision deal with special categories of data, as defined by applicable legal norms?
1.2.	May malicious parties have especially strong motives to hack the system? Can they easily achieve substantial financial gain - including by means of Blackmailing - or can a hacked system be used to achieve political goals (including expressing political opposition against the system)?

1.3.	Is the socio-technical system used to take, recommend, or affect decisions about individuals in a way that influences the outcome, i.e., what decision is taken?
1.4.	Is the system used to take a decision about a legal duty or right of an individual?
1.6.	Can individuals avoid the decision, or demand that the decision be taken via a different procedure, not involving the same technical system?
1.7.	Can the person about whom a decision is taken with the help of the relevant tool prove the wrongness of the decision without going to court?
1.8.	Is the harm of a wrong decision fully reversible?
1.9.	Is it possible to compensate the individual or family fully and adequately for the harm of a wrong decision, when it is ascertained that the decision was wrong and cannot be reversed?
1.10.	Does the decision concern any of the following areas of public life or public sector resources: the administration of justice, access to educational opportunities, access to democratic processes, etc.?
1.11.	Does the acquisition or the deployment of the ADMS result in a change in: – public computing infrastructure, – public data assets, or – intangible assets (e.g., competences) in the public sector?

1.12.	Is it possible that the technical system will have an effect on a political decision (e.g., a popular vote)?
1.13.	Does the technical system affect the distribution of public resources to economic actors in society?
1.14.	Does the technical system rely on a statistical model of human behavior or personal characteristics?
1.15.	Is the system designed to be adaptive so that it will not treat all new cases in the same way as those it encountered in the past, because it changes its parameters (e.g., in order to become more efficient)?
1.16.	Is the goal of the technical tool to automate a fully deterministic system of rules, which requires minimal creativity and human judgment by current human operators and which does not involve estimating risk or probability?
1.17.	Does the technical system rely on parameters, features, factors, or decision criteria that do not correspond to those normally considered by most experts in the field?
1.19.	Does the technical system rely on thirdparty infrastructure the public entity has no unrestricted control over and/or access to, e.g., data sets or computing power?

Se le risposte date alle domande della prima lista di controllo necessitano di approfondimento, vengono proposte ulteriori domande al fine di redigere dettagliatamente la relazione di valutazione.

4.2. Tool di valutazione dei sistemi di Intelligenza Artificiale

Nei capitoli precedenti abbiamo visto come negli ultimi anni sia cresciuta rapidamente l'attenzione verso le minacce e i rischi che possono sorgere con l'applicazione di sistemi di Intelligenza Artificiale. Proprio per questo, dopo aver svolto un periodo di analisi, ricerca e raccolta dei requisiti, abbiamo implementato uno strumento di valutazione dell'impatto dell'IA adatto al contesto italiano. Svolgere in anticipo una valutazione dell'impatto può aiutare nell'introduzione nella società di sistemi di IA affidabili contribuendo alla creazione di un ecosistema di fiducia verso queste tecnologie. Lo strumento aiuta, infatti, a comprendere gli aspetti legali ed etici di questa tecnologia tanto necessaria quanto pericolosa se progettata e utilizzata in maniera non adeguata. L'Intelligenza Artificiale sta svolgendo sempre più compiti e sta prendendo sempre più decisioni sia in completa autonomia sia in presenza degli umani. Molti sistemi, inoltre, hanno un grande impatto sulla società; di conseguenza, pensare a questi sistemi con un approccio all'etica è fondamentale. Questo strumento, che si basa sui quattro principi e i sette requisiti definiti dalla Commissione Europea, vuole quindi essere un supporto per sviluppatori, distributori e tutti coloro che desiderano utilizzare siste-

mi di IA a valutarne l'impatto su aspetti etici e legali e ad essere consapevoli in anticipo dei rischi che possono comportare (prevenendo errori che potrebbero essere in seguito costosi da risolvere).

Lo strumento ideato prevede quattro step principali: un **primo step** dedicato alla descrizione del sistema, un **secondo step** dedicato alla valutazione del rischio del sistema, un **terzo step** dedicato alla valutazione dell'impatto del sistema e, infine, un **quarto step** di restituzione dei risultati.

Dal primo al terzo step lo strumento è composto da alcune domande e diversi item al fine di indagare le aree definite nella seguente tabella:

Aree	Descrizione
PROGETTO	
Informazioni generali	Nome dell'organizzazione, titolo del progetto
Fase del progetto	Analisi – design – sviluppo – deployment – operation
Descrizione del progetto	Descrizione del sistema, dello scopo, del design, dei dati, del modello di addestramento, stakeholders
Benefici	Motivazione per l'introduzione del sistema

RISCHIO	
Profilo di rischio	Rilevatori ad alto rischio per i sistemi di IA
IMPATTO	
Intervento e sorveglianza umani	Autonomia umana, dignità umana e supervisione umana
Privacy e protezione dei dati	Protezione dei dati, privacy e buona governance
Robustezza tecnica e sicurezza	Sicurezza informatica, robustezza, accuratezza
Trasparenza	Tracciabilità (apertura su dati e algoritmi) e spiegabilità
Diversità, non discriminazione, equità	Equità, non discriminazione ed accessibilità
Benessere sociale e ambientale	Benessere ambientale, sociale e umano
Accountability	Accountability, tutela del diritto di contestazione

Il completamento dello strumento fornisce l'impatto o il rischio del sistema preso in esame su tre livelli: alto, moderato o basso (quarto step). In base alle risposte date, infatti, viene assegnato un determinato punteggio ad ogni area: in questo modo è possibile comprendere con facilità in quali aree sono presenti

criticità e di conseguenza in quali aree è necessario apportare modifiche affinché il sistema sia considerato affidabile. Lo strumento è uno dei pilastri del percorso di trasferimento tecnologico pensato per le aziende partecipanti al tavolo di lavoro "Artificial Intelligence Impact Assessment" all'interno della ReD OPEN Factory.

Postfazione

Donatella Paschina

"Tutti parlano di Intelligenza Artificiale (IA). Molti strumenti di IA sia hardware che software sono ora disponibili sul mercato e molti di più sono attesi quest'anno e nei prossimi. ***Cosa comporta questo per voi?****"*

Questa frase è tratta dall'introduzione ad un libro[1], divulgativo e non accademico, pubblicato nel 1987 [sic] negli Stati Uniti. Nonostante l'iperbole comunicativa sull'IA sia recente, i Sistemi basati su IA hanno già da anni fornito sostanziali vantaggi alle aziende che li hanno impiegati nei propri prodotti e servizi. Spesso però le organizzazioni e i clienti che li utilizzano quotidianamente non ne sono nemmeno pienamente consapevoli!

Aspettative e risultati

In una survey del 2017, condotta a livello globale, l'84% delle aziende intervistate dimostrava interesse per l'IA per il vantaggio competitivo che ne avrebbe tratto[2].

---

[1] Louis E. Frenzel, *Crash Course in Artificial Intelligence and Expert Systems*, 1987.

[2] S. Ransbotham, D. Kiron, P. Gerbert and M. Reeves, "Reshaping Business With Artificial Intelligence", MIT Sloan Management Review and The Boston Consulting Group, september 2017.

Nel 2021, un'analoga survey riporta che meno del 25% delle aziende intervistate che hanno utlizzato l'IA ha notato un impatto significativo sui risultati aziendali[3].

Nonostante la grande attenzione mediatica all'IA, l'applicazione si concentra spesso in campi specifici, in settori tecnologici avanzati o in soluzioni di automazione di processo.

Si parla quasi esclusivamente di IA rivolta a soluzioni per il riconoscimento visivo, il riconoscimento vocale, la guida autonoma, la robotica, la diagnostica avanzata, ma l'IA include una vastissima gamma di algoritmi e soluzioni a **supporto delle decisioni umane**, che non ambiscono a sostituire l'uomo, bensì ad **accrescerne le capacità cognitive e di soluzione di problemi complessi**.

Ma come assicurare un'implementazione di successo nella propria organizzazione?

Nell'assunto che una **strategia di adozione dell'IA** sia importante e non rimandabile per il futuro delle organizzazioni, è opportuno riflettere sugli **elementi** che costituiscono i **fattori di successo** dell'applicazione **dell'IA**, con specifico riferimento alle aziende non tecnologiche.

Consapevolezza e visione. Affrontare una strategia di adozione dell'IA richiede che il management sia informato e formato su cosa l'IA rappresenti per il futuro dell'azienda e per il settore industriale di appartenenza. Avere a bordo manager informati e consapevoli e ot-

[3] McKinsey Analytics, "Global Survey: The State of AI 2020", november 2021.

tenerne il commitment è infatti il principale fattore di successo per la formulazione e quindi l'esecuzione di un Piano IA.

Per acquisire le necessarie conoscenze e sviluppare un linguaggio comune è raccomandabile istituire un "laboratorio" in azienda composto dai manager, da esperti esterni e da esponenti accademici. Lo scopo è quello di analizzare quali aree siano indirizzabili verso una strategia IA nel settore di riferimento, venire a conoscenza di iniziative e progetti intrapresi da concorrenti o partner e quindi individuare quali pratiche siano già operative e, se conseguibili, quali risultati stiano portando.

Cultura e cambiamento. A fronte della formulazione di una strategia esplicita degli obiettivi raggiungibili attraverso l'adozione di strumenti basati su IA, occorre adoperarsi per comunicarla ai diversi livelli dell'organizzazione, in particolare con le risorse che saranno chiamate a eseguirla. Un modo efficace per disseminare la cultura IA in azienda e minimizzare eventuali resistenze dovute a timori o incomprensioni è avviare progetti focalizzati, dal perimetro ben definito, in determinate aree. Questo approccio richiede che le iniziative individuate nella fase di strategia vengano inquadrate in un Piano di adozione, prioritizzate per decidere su quale progetto investire e quindi attuate, stabilendo con chiarezza gli obiettivi e gli indicatori di performance che si vogliono raggiungere.

Persone. Ciascun Progetto di IA, seppure di dimensioni circoscritte, richiede un **team multifunzionale** composto da persone che posseggano le giuste com-

petenze, interne e esterne, e che ricoprano i ruoli fondamentali in un Progetto IA: Business/Process Owner, Data Analyst, esperti del settore di applicazione, specialisti IT e specialisti IA. Gli specialisti IA sono professionisti con competenze particolari, che difficilmente un'azienda non tecnologica riesce ad attrarre, formare con continuità e poi trattenere. È raccomandabile quindi, soprattutto all'inizio di un percorso di adozione dell'IA e in particolare per le PMI, una partnership con poli di trasferimento tecnologico piuttosto che assumere singoli individui.

Modelli. Gli algoritmi e le tecniche di IA sono sempre più disponibili sul mercato dell'offerta di soluzioni tecnologiche. Occorre saperli scegliere e utilizzare. Possono essere "noleggiati" nei data center dei provider tecnologici, accorciando fortemente i tempi di disponibilità della soluzione, concentrando lo sforzo progettuale non nello sviluppo di tecniche complesse, ma nel configurare gli algoritmi alle peculiarità del processo oggetto di applicazione e alimentarli con i dati necessari.

Dati. Gli algoritmi di IA necessitano di grandi quantità di dati per funzionare e produrre soluzioni o alternative di soluzione. I dati necessari sono sia quelli prodotti dai sistemi interni all'azienda sia eventualmente dati esterni. Questi ultimi possono essere acquisiti da fonti esterne, mentre quelli interni devono essere resi disponibili per alimentare il modello/sistema basato su tecniche IA. Spesso ciò comporta la necessità di un "progetto preliminare", talvolta trascurato in termini di costi e tempi, volto proprio a rendere disponibili le

basi dati su cui poi applicare gli algoritmi selezionati.

Il 2022 segna un nuovo corso per l'IA:

- **L'Intelligenza Artificiale non è più appannaggio esclusivo delle aziende tecnologiche.** Ogni organizzazione, aziendale o pubblica, può introdurre soluzioni e strumenti a vantaggio dei risultati e del miglioramento dei propri prodotti e servizi.
- **Il Programma Strategico Intelligenza Artificiale 2022-2024 del Governo Italiano**[4] ha indicato i Principi, gli Obiettivi e gli Strumenti finanziari a sostegno della Strategia Nazionale per l'IA.
- I manager e gli amministratori si adopereranno affinché **la Strategia IA diventi parte integrante della Strategia dell'organizzazione**, quale fattore di successo abilitante il raggiungimento degli obiettivi di modernizzazione e di creazione di nuovo valore.
- **Il Piano di Trasformazione Digitale** delle aziende e **l'Agenda Digitale delle amministrazioni pubbliche includeranno** anche le iniziative e **i Progetti di IA.**

[4] innovazione.gov.it/notizie/articoli/intelligenza-artificiale-l-italia-lancia-la-strategia-nazionale/, 24 Novembre 2021.

ReD OPEN FACTORY - Center for Responsible Innovation è il Centro per l'innovazione responsabile creato da ReD OPEN per la diffusione, co-creazione e co-progettazione di **regole, modelli e metodologie operative di una data strategy** consapevolmente responsabile.

Il Center for Responsible Innovation:

- ha la missione di **diffondere nelle imprese la ricerca** non ancora applicata e di rappresentare la forma permanente di un modello di business orientato al trasferimento tecnologico tra ricerca e impresa;
- media, per conto delle imprese, **linguaggi e processi potenzialmente innovativi** ma ancora lontani dall'essere "mainstream";
- si fonda su **logiche partecipative e condivise**, atte a trasferire alle aziende modelli e piattaforme operative di governance dell'innovazione che abilitano la trasformazione digitale in modo «consapevole».

ReD OPEN è uno spin-off dell'**Università di Milano-Bicocca**, nato con l'intento di accompagnare e aiutare le imprese per affrontare percorsi di innovazione, transizione digitale e ricorso all'Intelligenza Artificiale in modo «responsible by design». Con competenze ed esperienze multidisciplinari, ReD OPEN propone percorsi di accompagnamento e di riconfigurazione dei modelli organizzativi e di business, in chiave responsabile.

Per approfondimenti:
www.redopenletter.it
www.redopenfactory.com
www.redopen.it

Il volume intende facilitare la conoscenza e l'approfondimento di un tema che sta tuttora alimentando discussioni e conversazioni nei contesti giuridici, governativi e industriali sia in Italia che all'estero: l'Intelligenza Artificiale e il suo uso.
È fondamentale avere consapevolezza dell'impatto che l'Intelligenza Artificiale ha e potrà avere nella vita quotidiana dal punto di vista della trasformazione digitale già in corso da anni, con conseguenze sulla privacy dell'individuo, sugli iter normativi nazionali e internazionali e sul governo delle tecniche decisionali; proprio per questo motivo il lavoro, pensato anche per i lettori meno esperti, è utile per chi vuole comprendere i principi fondamentali dell'IA e di conseguenza poter seguire e approfondire le evoluzioni in corso.

www.ledizioni.it

9 788855 266222

€ 12,90