IL REGOLAMENTO EUROPEO SULL'INTELLIGENZA ARTIFICIALE

Analisi informatico-giuridica

GIUSEPPE CONTISSA
FEDERICO GALLI
FRANCESCO GODANO
GALILEO SARTOR

i-lex

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista semestrale on-line: www.i-lex.it
Dicembre 2021
Fascicolo 2
ISSN 1825-1927

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista semestrale on-line: www.i-lex.it

IL REGOLAMENTO EUROPEO SULL'INTELLIGENZA ARTIFICIALE
ANALISI INFORMATICO-GIURIDICA

GIUSEPPE CONTISSA*
FEDERICO GALLI
FRANCESCO GODANO
GALILEO SARTOR

Abstract. Il 21 aprile 2021 la Commissione europea ha pubblicato la proposta di regolamento che stabilisce norme armonizzate sull'intelligenza artificiale (Artificial Intelligence Act). La proposta fornisce una definizione giuridica di IA, una metodologia per definire i sistemi IA ad alto rischio e dettaglia un quadro normativo al fine di fornire un quadro normativo per i sistemi IA compatibile con i diritti umani e i valori europei. Le regole riguardano i requisiti di progettazione e di gestione per i sistemi IA ad alto rischio, gli obblighi specifici per i fornitori di IA e altri attori coinvolti, i requisiti di trasparenza per i sistemi IA specifici, e un quadro complesso per l'applicazione, che comprende la standardizzazione, la certificazione e l'autovalutazione, nonché un nuovo sistema di supervisione, incentrato sull'Artificial Intelligence European Board. Il presente contributo analizza le principali questioni giuridiche e mette in luce l'approccio regolatorio. Trae alcune conclusioni per le raccomandazioni politiche.

Parole chiave: *Artificial Intelligence Act, pratiche proibite, definizione di IA, sistemi IA ad alto rischio, governance* (max. 6)

* Giuseppe Contissa (giuseppe.contissa@unibo.it), ALMA AI-CIRSFID Alma Mater Studiorum Università di Bologna

Federico Galli (federico.galli7@unibo.it), ALMA AI-CIRSFID Alma Mater Studiorum Università di Bologna

Francesco Godano (francesco.godano@unibo.it), ALMA AI-CIRSFID Alma Mater Studiorum Università di Bologna

Galileo Sartor (galileo.sartor2@unibo.it), ALMA AI-CIRSFID Alma Mater Studiorum Università di Bologna

1 Introduzione

Il presente contributo esamina la recente Proposta di Regolamento sull'Intelligenza artificiale, presentata dalla Commissione europea nell'aprile 2021. La Proposta mira a costruire un quadro normativo dell'Intelligenza artificiale (IA) compatibile con i valori europei e con la tutela dei diritti umani: fornisce una definizione e classificazione dei sistemi di IA, stabilisce una serie di requisiti per la loro realizzazione, commercializzazione e utilizzo, e delinea una serie di strumenti di controllo e di sanzioni. In questa sede descriveremo i tratti salienti della Proposta legislativa, vagliandone gli aspetti innovativi, gli elementi problematici ed i possibili interventi di modifica.

Negli ultimi anni il campo dell'IA ha conosciuto un rapido sviluppo in settori sempre più ampi della vita sociale. Grazie alle capacità di diffusione di Internet, le tecnologie dell'informazione e di IA stanno trovando crescente applicazione non solo nei ristretti ad alta specializzazione tecnologica, ma anche all'interno di prodotti e servizi di uso comune, coinvolgendo ampie fasce della popolazione. A fronte degli enormi benefici, molti sono i rischi che tali tecnologie comportano nei confronti dei diritti e delle libertà fondamentali. Si è quindi posta all'attenzione delle istituzioni europee la necessità di fornire una base normativa comune che consideri le implicazioni umane, etiche e socio-economiche dell'IA nel rispetto dei principi e dei valori europei. A partire dal 2017, una serie di iniziative sono state avviate per elaborare i principi, la struttura e i contenuti di tale legislazione, nel contesto di un più generale programma che coinvolge la regolazione della gestione dei dati, dei servizi del mercato digitale e della robotica[1].

[1] Di recente emanazione sono: Commissione europea, Proposta di Regolamento del Parlamento europeo e del Consiglio relativo a mercati equi e contendibili nel settore digitale (legge sui mercati digitali), 15 dicembre 2020, COM(2020) 842 final (Digital Markets Act); Commissione europea, Proposta di Regolamento del Parlamento europeo e del Consiglio relativo a un mercato unico dei servizi digitali (legge sui servizi digitali) e che modifica la direttiva 2000/31/CE, 15 dicembre 2020, COM(2020) 825 final (Digital Services Act); Proposta di Regolamento del Parlamento europeo e del Consiglio relativo alla governance europea dei dati (Atto sulla governance dei dati), 25 novembre 2020, COM(2020) 767 final (Data Governance Act).

I primi documenti in cui le istituzioni europee sollevano la necessità di aggiornare il quadro giuridico comunitario risalgono al 2017[2]. Nella Comunicazione del 25 aprile 2018, la Commissione Europea ha presentato una "strategia europea" in tema di IA. La strategia è orientata su tre direttrici: (1) la promozione della ricerca, dello sviluppo tecnologico e delle applicazioni industriali legate all'IA; (2) il sostegno all'apparato socio-economico interessato da tali sviluppi tecnologici; (3) infine, la definizione di "un quadro etico e giuridico adeguato, basato sui valori dell'Unione e coerente con la Carta dei diritti fondamentali dell'UE"[3].

Parallelamente, la Commissione ha istituito un "Gruppo di esperti di alto livello" sull'IA che ha pubblicato, nell'aprile 2019, gli *Orientamenti etici per un'Intelligenza artificiale affidabile*[4]. Tali Orientamenti individuano tre componenti fondamentali per un'IA che possa considerarsi "degna di fiducia" (*trustworthy*): legalità, eticità, robustezza tecnica e sociale. Sulla base di tali principi, il Gruppo di esperti ha individuato sette requisiti per un'IA affidabile[5]: (1) intervento e sorveglianza umani; (2) robustezza tecnica e sicurezza; (3) riservatezza e governance dei dati; (4) trasparenza; (5) diversità, non

[2] Risoluzione del Parlamento europeo del 16 febbraio 2017 recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica, 2015/2103(INL); Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni sulla revisione intermedia dell'attuazione della strategia per il mercato unico digitale. Un mercato unico digitale connesso per tutti, 10 maggio 2017, COM(2017) 228 final.

[3] Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato economico e sociale Europeo e al Comitato delle regioni. L'Intelligenza artificiale per l'Europa, 25 aprile 2018, COM(2018) 237 final, pp. 3-4. Oltre alla Carta dei diritti, la Commissione si riferisce all'articolo 2 del TUE per la definizione dei valori fondamentali dell'Unione.

[4] Gruppo di esperti ad alto livello sull'Intelligenza artificiale, Orientamenti etici per un'IA affidabile, 8 aprile 2019.

[5] Le linee guida contengono inoltre una "lista di controllo" per la valutazione dell'operatività di tali requisiti da parte delle aziende. Durante la seconda metà del 2019, oltre 350 organizzazioni hanno sottoposto i propri prodotti e servizi di IA al vaglio di questa lista: sulla base dei riscontri ricevuti è emerso che, mentre un certo numero di requisiti si riflette già nei regimi normativi esistenti, quelli riguardanti la trasparenza, la tracciabilità e la supervisione umana non sono specificamente coperti dalla legislazione vigente in molti settori economici.

discriminazione ed equità; (6) benessere sociale e ambientale; (7) *accountability*.

Sulla base di tale quadro etico-giuridico, lo stesso Gruppo di esperti ha pubblicato anche una raccomandazione sulle policy e sugli investimenti in materia di IA[6]. Raccogliendo e sviluppando le suddette indicazioni, nel febbraio 2020 la Commissione Europea ha pubblicato il *Libro bianco sull'Intelligenza artificiale*[7]. Il documento articola in modo dettagliato le opzioni strategiche definite nel 2018 per un'IA "made in Europe", affiancando all'obiettivo di una tecnologia "degna di fiducia" la promozione dell'"eccellenza" europea nel settore della ricerca e dell'industria. Per quanto riguarda l'ambito giuridico, il Libro bianco definisce come prioritari gli interventi in tema di diritti fondamentali e governance dei dati, di sicurezza e regimi di responsabilità, di armonizzazione ed effettività della legislazione.

Dal canto suo, il Parlamento europeo ha anch'esso affrontato, in una serie di risoluzioni, alcune questioni sollevate dalla diffusione delle tecnologie di IA. In particolare, la Risoluzione sulla responsabilità[8] è rilevante per la definizione di IA contenuta nell'art. 3(a) della Proposta, e per l'adozione di un approccio alla regolamentazione basato sul rischio.

Il percorso, qui rapidamente sintetizzato, è culminato nella Proposta di un Regolamento Europeo, noto come "AI Act"[9], che predispone un quadro normativo organico per lo sviluppo e l'utilizzo dell'IA.

---

[6] High-Level Expert Group on Artificial Intelligence, Policy and Investment Recommendations for Trustworthy AI, 26 Giugno 2019

[7] Commissione Europea, Libro Bianco sull'Intelligenza artificiale. Un approccio europeo all'eccellenza e alla fiducia, 19 febbraio 2020, COM/2020/65 final.

[8] Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione su un regime di responsabilità civile per l'intelligenza artificiale, 2020/2014(INL).

[9] Commissione Europea, Proposta di Regolamento del Parlamento Europeo e del Consiglio che stabilisce regole armonizzate sull'Intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione (AI Act), 21 aprile 2021, COM/2021/206 final.

2 La Proposta di Regolamento: caratteri essenziali e ambito di applicazione

Andiamo ora ad esaminare la struttura e gli elementi principali della Proposta. Il documento è articolato in 12 titoli, che ricomprendono 85 articoli, preceduti da 89 considerando. Il testo è corredato da 9 allegati tecnici.

La prima parte della Proposta, che ricomprende il Titolo I, definisce l'oggetto e il campo di applicazione del Regolamento. Le nuove norme sono dirette: (1) ai fornitori che immettono sul mercato o mettono in servizio sistemi di IA all'interno dell'Unione; (2) agli utenti di sistemi di IA situati nell'Unione; (3) ad utenti e fornitori di sistemi situati in un paese terzo, quando "l'output prodotto dal sistema sia utilizzato nell'Unione" (art. 2).

L'art. 3 contiene le definizioni utilizzate all'interno del documento. Di primaria importanza, fra di esse, è la definizione di "sistema di intelligenza artificiale", di cui si discuterà dettagliatamente nella Sezione 3 del presente contributo. In materia, è opportuno sin da subito segnalare l'utilizzo da parte della Proposta del temine "utente". Con questa dizione, ci si riferisce a "qualsiasi persona fisica o giuridica […] che utilizza un sistema di IA sotto la sua autorità": la nozione differisce dunque da quella comune di "utilizzatore finale" del sistema (vedi Sezione 6).

La seconda parte della Proposta (titoli II-IV) contiene la disciplina delle diverse categorie di sistemi di IA individuate dal legislatore europeo. La classificazione prescelta segue un approccio cd. *risk-based*, fondato sul livello di rischio che i caratteri o l'uso di tali tecnologie comportano. Si individuano in tal modo i sistemi che comportano (i) un rischio inaccettabile, (ii) un rischio alto, e (iii) un rischio medio o basso.

Il titolo II, che coincide con il solo art. 5, disciplina le "pratiche" – ossia gli utilizzi o gli effetti dell'IA – qualificate come proibite nell'ordinamento comunitario, in ragione del loro patente contrasto con i valori o i diritti fondamentali dell'Unione. Il divieto riguarda quattro categorie: le pratiche cd. di manipolazione, che (1) utilizzano tecniche subliminali o (2) altrimenti sfruttano la vulnerabilità di determinati gruppi di persone al fine di distorcerne il comportamento in modo potenzialmente dannoso; (3) le pratiche cd. di *social scoring*, che

utilizzano l'IA allo scopo di valutare o classificare l'affidabilità delle persone in determinati contesti sociali (ma solo, si badi, da parte di autorità pubbliche); infine, (4) l'uso di sistemi di identificazione biometrica a distanza (riconoscimento facciale e vocale) in tempo reale in spazi pubblicamente accessibili e per fini di *law enforcement*. Ci soffermeremo su ognuna di queste pratiche nella Sezione 4.

Il titolo III contiene la regolamentazione dei sistemi di IA che creano un rischio elevato per la salute e la sicurezza o i diritti fondamentali delle persone fisiche. Questi sistemi, di cui parleremo nella Sezione 4, sono ammessi nell'ordinamento dell'Unione nella misura in cui soddisfino determinati requisiti e siano sottoposti ad una valutazione di conformità preliminare alla loro immissione sul mercato o messa in servizio.

I sistemi "ad alto rischio" sono individuati, al Capo 1 del Titolo III, con riferimento a due criteri diversi: la destinazione d'uso come componenti di sicurezza di determinate categorie di prodotti; la presenza di funzionalità o l'appartenenza a determinati settori socio-economici che incidono sui diritti fondamentali. Il Capo 2 stabilisce i requisiti obbligatori di tali sistemi, con particolare riferimento alla gestione dei dati, agli obblighi di trasparenza e spiegabilità (*explainability*) del sistema, alla sicurezza e robustezza tecnica, alla supervisione umana. Il Capo 3 impone una serie di obblighi ai fornitori, agli utenti e agli altri soggetti coinvolti nel ciclo di vita dei sistemi.

I Capi 4 e 5 disciplinano invece gli organismi e le procedure relative alla valutazione di conformità *ex ante* cui tali sistemi sono soggetti, basato su un modello di controllo interno alle organizzazioni (e che si completa con i meccanismi di controllo esterni *ex post*, di cui ai Titoli VI e seguenti). La Sezione 5 descrive analiticamente le norme relative a questi sistemi di controllo.

Il Titolo IV della Proposta dispone alcuni obblighi di trasparenza – diversi e ulteriori rispetto a quelli previsti per i sistemi ad alto rischio – in capo a determinati sistemi di IA che presentano specifici rischi per le persone fisiche. Si tratta di quei sistemi che (i) interagiscono con gli esseri umani; oppure (ii) sono utilizzati per rilevare le emozioni o determinare l'associazione con categorie (sociali) sulla base di dati biometrici; o ancora che (iii) generano o manipolano contenuti audio-visivi ("deep fakes"). In tali casi, si prevede che chi utilizza il sistema debba essere informato del fatto che sta interagendo con un'IA. I

sistemi elencati corrispondono generalmente ad una tipologia di IA che più sopra abbiamo definito "a medio rischio", e che in tal modo viene infatti spesso classificata. Tuttavia, occorre notare che anche i sistemi ad alto rischio possono ricadere in questa casistica, e dunque essere soggetti al relativo obbligo di trasparenza. Di questi argomenti si occuperà nel dettaglio la Sezione 6.**Error! Reference source not found.**

La terza parte della Proposta (Titoli V-XII) contiene la normativa sulla governance e sul controllo dei sistemi di IA, insieme alle regole sull'esecuzione del Regolamento. Dei tratti fondamentali di questa parte ci occuperemo, più sinteticamente, nella Sezione 5.

Il Titolo V si pone l'obiettivo di creare un quadro giuridico favorevole all'innovazione tecnologica e sociale. A tale fine, si promuove la creazione di "ambienti normativi controllati" (*normative sandboxes*) per la sperimentazione tecnologica, fornendone una regolamentazione essenziale.

Il Titolo VI disciplina gli organismi di governo del sistema-mercato di IA a livello dell'Unione e degli Stati membri. A tal fine, il legislatore europeo prevede l'istituzione di un Comitato Europeo per l'Intelligenza artificiale, con compiti di coordinamento e supervisione del sistema. Si fornisce inoltre una disciplina per le autorità nazionali competenti in materia.

Il Titolo VII mira ad agevolare le funzioni di governance attraverso la creazione di una banca dati europea per i sistemi di IA ad alto rischio.

Il Titolo VIII fornisce la regolamentazione del monitoraggio successivo all'immissione sul mercato dei sistemi di IA, degli obblighi informativi e della vigilanza del mercato.

Nel Titolo IX si fornisce una serie di norme relative alla creazione di codici di condotta per i fornitori di sistemi di IA non ad alto rischio, che mirano a incoraggiare l'adozione volontaria, anche da parte di questi ultimi, dei requisiti obbligatori per i sistemi ad alto rischio.

Gli ultimi titoli contengono norme di carattere generale relative all'esecuzione del futuro Regolamento: l'obbligo di riservatezza nella gestione delle informazioni, disposto in capo alle autorità pubbliche (Titolo X); le norme relative al potere di adottare atti delegati (Titolo XI); gli obblighi per la Commissione di valutare regolarmente la necessità di aggiornamenti e di preparare relazioni periodiche sulla valutazione e la revisione del Regolamento (Titolo XII).

Entriamo ora nel dettaglio delle norme.

3 Definizione di Intelligenza artificiale

La definizione di Intelligenza artificiale inclusa nella Proposta all'art. 3, n.1 è composta da due parti. La prima, funzionale, intende essere più ampia possibile ed include il software che "può, per una determinata serie di obiettivi definiti dall'uomo, generare output quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono".

La seconda è costituita da un elenco dettagliato di tecniche e approcci che definiscono il perimetro dei sistemi di IA contenuti nell'Allegato I. Si fa riferimento a tre modelli generali: apprendimento automatico (apprendimento supervisionato e non, apprendimento per rinforzo), approcci basati su logica e modelli espliciti della conoscenza (come i sistemi esperti), e approcci statistici (come la stima bayesiana)[10].

Le due parti della definizione tendono evidentemente in direzioni diverse. La definizione generale, più in linea con il modello *risk-based* adottato dalla Proposta, è neutrale rispetto alla tecnologia e guarda alle capacità di influenza del sistema sul mondo esterno. Al contrario, l'elenco delle tecniche nell'Allegato I cerca di restringere il cerchio di applicazione a tecnologie specifiche, evidentemente rispondendo all'esigenza di dare maggiore certezza ai produttori e utilizzatori di sistemi di IA.

La scelta di includere tale lista suscita diverse perplessità, in particolare in relazione all'ambito dei sistemi coperti dalla Proposta, e in relazione all'approccio basato sul rischio che il legislatore stesso dichiara di voler adottare.

Il riferimento ai tre approcci di IA potrebbe condurre a una interpretazione molto ampia, tale da coprire tecnologie che normalmente non sono considerate di IA. In particolare, il riferimento a modelli logici e statistici rischia di estendere molto il campo di applicazione del Regolamento, sostanzialmente includendo tutti o

[10] Nell'articolo 4 della Proposta si stabilisce l'uso di atti delegati per modificare l'elenco delle tecniche e degli approcci definiti nell'allegato I per mantenere l'elenco aggiornato agli sviluppi futuri del mercato e delle tecnologie.

quasi i sistemi di decisione algoritmica.[11] In questa prospettiva, verrebbe da chiedersi quale sia la ragione di definire specifici modelli di IA, e ancor prima, di focalizzare la Proposta sulla regolazione dell'IA, e non sulle decisioni algoritmiche in senso lato. Si deve inoltre rilevare che il confine fra le varie tecniche elencate non è sempre chiaro (ad esempio tra calcoli complessi e approcci statistici, o ancora tra approcci logici e procedurali). Il rischio è quello di generare notevole incertezza, anche tra gli operatori del mercato, al momento di determinare se un certo sistema di IA adotti o meno le tecniche definite nell'Allegato I.

D'altro canto, se la definizione di IA fosse interpretata in senso stretto, potrebbe finire per escludere alcune applicazioni di sistemi automatizzati che, sebbene sviluppati attraverso metodi tradizionali (come la programmazione procedurale) e completamente deterministici, presentano un alto grado di rischio, specialmente quando progettati per essere indipendenti dall'intervento/supervisione umana. In questo gruppo potremmo includere sistemi come Frank[12], che assegnava le prenotazioni ai *rider* di Deliveroo, e il sistema Parcoursup per l'iscrizione all'università usato in Francia[13][14]. Entrambi questi sistemi si basavano su complesse forme di automazione, non necessariamente riconducibili alle tecniche di IA previste dalla Proposta, e sono stati comunque oggetto di controversie legali e al centro del dibattito giuridico per il loro impatto profondo e invasivo sui diritti e le libertà degli individui.

---

[11] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way*, Springer Nature, 2019.

[12] F. Galli, F. Godano, *Il rapporto di lavoro dei riders e la natura discriminatoria delle condizioni di accesso al lavoro dell'algoritmo frank*, Diritto di Internet, III, 2021, pp. 275–288.

[13] A. Bibal, M. Lognoul, A. De Streel, B. Frénay, *Legal requirements on explainability in machine learning*, Artificial Intelligence and Law, 29 (2021), pp. 149–169.

[14] Si è discusso come il sistema abbia avuto un profondo impatto sulla vita degli studenti, determinando quali posti di laurea sarebbero disponibili per loro dopo la scuola superiore. Mentre il funzionamento del sistema non è mai stato descritto pubblicamente, sembra che non siano utilizzate né tecniche statistiche, né di apprendimento automatico, né basate sulla conoscenza, piuttosto una combinazione di più metodi, con un grado di supervisione umana nella decisione.

L'approccio "*cherry-picking*" nella definizione dell'IA porta perciò ad escludere dal Regolamento alcuni sistemi che presentano un elevato livello di rischio, comparabili a quelli dei sistemi inclusi, e d'altra parte ad includere sistemi che, sebbene facciano ricorso alle tecniche elencate, non necessariamente presentano profili di rischio. Rimangono dubbi su alcuni punti. Non è chiaro come potrebbero essere definite altre tipologie di sistemi (come i sistemi multi-agente), o se le considerazioni del Regolamento si possano applicare allo stesso modo a un programma basato su regole di molto risalente nel tempo e a un complesso sistema di apprendimento automatico sviluppato oggi[15]. Tutto ciò sembra contrastare con l'obiettivo, dichiarato dal legislatore stesso, di voler utilizzare un approccio basato sul rischio e tecnologicamente neutrale.

Riassumendo, la Proposta di Regolamento adotta un sistema doppio di definizione per l'Intelligenza artificiale, mirando ad includere i software che fanno uso di sistemi di *machine learning*, programmazione logica, o metodi statistici. Mentre la prima categoria sarebbe probabilmente troppo limitata, le altre due includono sistemi di ricerca e ottimizzazione che classificherebbero erroneamente come sistemi di IA la grande maggioranza di software esistenti[16].

In base a questa definizione, una prima considerazione riguarda il fatto che la Proposta di Regolamento avrebbe potuto parlare non di IA, bensì del rischio delle decisioni algoritmiche: ambito più generale ma non di minore importanza, che avrebbe potuto includere i sistemi identificati dall'attuale Proposta, senza fare distinzioni tecnologiche.

Vale la pena ricordare che nella già citata Risoluzione del Parlamento europeo su un regime di responsabilità civile per l'Intelligenza artificiale 2020/2014(INL) è stata adottata una definizione molto più generica e astratta, ovvero "un sistema che è basato su software o incorporato in dispositivi hardware, e che mostra un comportamento che simula l'intelligenza, tra l'altro, raccogliendo ed elaborando dati, analizzando e interpretando il suo ambiente, e intraprendendo azioni, con un certo grado di autonomia, per

[15] La circostanza è stata rilevata anche da Barry O´Sullivan, membro del Gruppo di esperti ad alto livello sull'Intelligenza artificiale. Si veda il suo intervento accessibile online: https://www.youtube.com/watch?v=X9h8MZIuvKg.

[16] P. Glauner, *An Assessment of the AI Regulation Proposed by the European Commission*, 2021.

raggiungere obiettivi specifici". Questa definizione, peraltro più simile a quella proposta originariamente dal Gruppo di esperti di alto livello, appare più neutrale dal punto di vista tecnologico e sembrerebbe più in linea con una valutazione dell'IA basata sul rischio.

4 Le pratiche proibite

Il Titolo II della Proposta, contenente il solo art. 5, circoscrive i sistemi di IA caratterizzati dal più elevato livello di rischio. Qui il legislatore europeo adotta una tecnica di classificazione leggermente diversa: non si occupa di specifici sistemi di IA e del loro uso, ma di alcune pratiche realizzate attraverso sistemi di IA che generano o possono generare effetti considerati inaccettabili poiché in contrasto con i valori e i diritti fondamentali dell'Unione. Per queste pratiche, la Proposta pone un divieto generalizzato, salvo ammettere eccezioni in casi particolari. Di seguito analizziamo le quattro pratiche proibite

4.1 Manipolazione

L'articolo 5 (1), lett. a, proibisce "l'immissione sul mercato, la messa in servizio o l'uso di un sistema di IA che utilizza tecniche subliminali che agiscono senza che una persona ne sia consapevole al fine di distorcerne materialmente il comportamento in un modo che provochi o possa provocare a tale persona o a un'altra persona un danno fisico o psicologico".

Rispetto a una versione precedente della Proposta, le forme di manipolazione vietate appaiono più dettagliate, sia nella loro dimensione strutturale, sia con riguardo agli effetti. Per la prima dimensione, la manipolazione non è più definita in relazione all'architettura e all'interfaccia digitale[17], ma è sostituita da un'azione specifica: il dispiegamento di tecniche subliminali al di là della consapevolezza di una persona. Quanto agli effetti, il divieto dell'art.

[17] L'art. 4 della versione non ufficiale recitava "AI systems designed or used in a manner that manipulates human behaviour, opinions or decisions through choice architectures or other elements of user interfaces, causing a person to behave, form an opinion or take a decision to their detriment."

5 include solamente la distorsione del comportamento di una persona che causa (anche solo potenzialmente) danni fisici o psicologici. La versione precedente contemplava, più in generale, tutte le possibili attività volte ad influenzare i comportamenti degli individui (incluso le opinioni e decisioni) che potessero arrecare un pregiudizio a questi ultimi.

Per quanto più specifica nella sua formulazione, la nuova versione del divieto lascia aperte alcune importanti questioni. Il significato dell'espressione "tecniche subliminali" non è chiaro, né appare comprensibile il significato di "consapevolezza" in relazione alle pratiche che opererebbero al di là di essa. Nella versione del testo in inglese, è utilizzato il termine "*consciousness*" che sembra conferire al divieto un significato metagiuridico e filosofico. Ci si potrebbe chiedere se, alla luce di questa vaga terminologia, le tecniche che erano esplicitamente vietate nella versione precedente siano ancora incluse. Consideriamo ad esempio l'utilizzo delle architetture digitali per indurre certi comportamenti nocivi negli utenti, i cd. *dark pattern*[18]. Si pensi, inoltre, al cd. *hyper-nudging*, ossia quelle tecniche adattive e personalizzate di analisi di dati che generano nell'individuo comportamenti irrazionali ed impulsivi[19].

Relativamente al risultato della manipolazione, le tipologie di danno riferite nel divieto non sono chiare. Sul punto la Commissione avrebbe potuto essere più esplicita: mentre il danno fisico in teoria è facile da identificare, appare più difficile immaginare casi in cui il danno fisico sia provocato da sistemi di IA usati per manipolare le persone[20]. Viceversa, gli scenari in cui potrebbero verificarsi danni psicologici appaiono più verosimili considerato il crescente utilizzo di

[18] Luguri, L. Strahilevitz, L., *Shining a light on dark patterns*, University of Chicago, Public Law Working Paper, 2019.

[19] Yeung, K., *'Hypernudge': Big Data as a mode of regulation by design*, Information, Communication &; Society, 20, 2017, pp. 118-136.

[20] La Commissione ha presentato solo un esempio piuttosto inverosimile: "[un] suono impercettibile [riprodotto] nelle cabine dei camionisti per spingerli a guidare più a lungo di quanto sia sano e sicuro [dove] l'IA è usata per trovare la frequenza che massimizza questo effetto sui conducenti". Vedi, Gabriele Mazzini (DG CONNECT), 'A European Strategy for Artificial Intelligence' (2nd ELLIS Workshop in Human-Centric Machine Learning (YouTube recording), 10 Maggio 2021) https://youtu.be/OZtuVKW qhl0?t=10346.

sistemi intelligenti capaci di riconoscere gli stati emotivi degli individui[21] e di influire su di essi:[22] tuttavia, in assenza di chiari indici e mezzi di misurazione, risulta difficile identificare il danno psicologico, vista la sua intrinseca soggettività. La vaghezza della formulazione potrebbe portare a pensare che anche la pubblicità mirata sulle piattaforme digitali possa comportare danni psicologici, ogniqualvolta la persona provi fastidio o irritazione per l'intrusione nella propria sfera privata. Occorre, infine, segnalare l'esclusione dei danni economici e dei pregiudizi di carattere sociale, scelta che escluderebbe molte pratiche di IA ritenute rischiose. Tra queste, l'utilizzo di sistemi di IA volti ad orientare i comportamenti dei consumatori relativamente alle scelte economiche (ad es., sistemi per la pubblicità mirata o sistemi di raccomandazione) o le preferenze politiche[23].

In ultima analisi, il rischio è che, anche se ben intenzionato nel suo obiettivo, il divieto di manipolazione condotta attraverso sistemi di IA rimanga una dichiarazione di intenti senza alcuna applicazione concreta. I pericoli di manipolazione legati all'uso dell'IA che sono stati discussi negli ultimi anni sono stati riconosciuti dalla Proposta, e questo merita una valutazione positiva. Tuttavia, ci si augura che il legislatore vada ben oltre un vago riferimento al tema della manipolazione (visibile nel riferimento alla nozione di "tecniche subliminali", concetto risalente agli albori della pubblicità e del marketing di massa)[24] e specifichi quali pratiche o tecniche siano da considerare così pericolose da mettere in pericolo anche a lungo termine l'autonomia degli individui. Per esempio, ci si potrebbe

[21] Burr, C., Cristianini, N., *Can Machines Read our Minds?*, Minds and Machines, (2019), pp. 1-34.

[22] Matz, S. C., Kosinski, M., Nave, G., Stillwell, D. J., *Psychological targeting as an effective approach to digital mass persuasion*, Proceedings of the National Academy of sciences, 114, 2017, pp. 12714-12719.

[23] Ci sembra importante menzionare che, in caso di danno economico, si applica comunque la Direttiva 2005/29/CE sulle pratiche commerciali scorrette. La Direttiva vieta, tra le altre cose, le pratiche che, contrariamente alle norme di diligenza professionale, falsano o sono idonee a falsare in misura rilevante il comportamento economico del consumatore medio. Sull'applicabilità della direttiva alle pratiche scorrette poste in essere attraverso sistema di IA, si veda BEUC, *EU Consumer protection 2.0. Structural asymmetries in digital consumer market*, Report, March 2021.

[24] Packard, V., Payne, R., *The hidden persuaders*, McKay New York, 1957.

aspettare un divieto dell'utilizzo di certe tecniche di apprendimento automatico (come l'apprendimento per rinforzo) in contesti di più facile dipendenza come i videogiochi, i mercati finanziari, o i social network. Allo stesso modo, si potrebbe pensare di vietare l'uso di alcune categorie di dati sensibili a seconda del contesto, data la loro capacità di influenzare l'autonomia di scelta e i comportamenti (es. dati biometrici in contesti di marketing, informazioni sulle opinioni politiche nei social network).

4.2 Sfruttamento di gruppi vulnerabili

La seconda pratica proibita consiste nell'uso di un sistema "che sfrutta le vulnerabilità di uno specifico gruppo di persone, dovute all'età o alla disabilità fisica o mentale, al fine di distorcere materialmente il comportamento di una persona che appartiene a tale gruppo in un modo che provochi o possa provocare a tale persona o a un'altra persona un danno fisico o psicologico" (art. 5, lett. b). Il divieto sembrerebbe aggiungere un ulteriore strato di protezione contro la manipolazione nei casi in cui sono coinvolti gruppi specifici considerati vulnerabili.

L'attenzione ai gruppi più deboli nella Proposta è lodevole. Allo stesso tempo, però, non è chiaro quale sia l'effettivo valore aggiunto di questa disposizione rispetto alla precedente. In particolare, non viene dato alcun indizio sul significato di "sfruttamento". Si potrebbe pensare che lo sfruttamento sia diverso dalla manipolazione, e che l'uso delle informazioni possa qualificare alcune persone fisiche come vulnerabili. Per esempio, un sistema che utilizza le informazioni sull'età di una persona al fine di promuovere l'acquisto di articoli per bambini sembra essere coperto dalla disposizione.

Inoltre, per quanto riguarda l'approccio concettuale alla vulnerabilità, si fa riferimento a gruppi che, per caratteristiche specifiche come l'età e la disabilità fisica o mentale, sono più tradizionalmente vulnerabili[25]. Non è chiaro perché questi e non altri

[25] Anche in questo divieto, il Regolamento sembra seguire pedissequamente la Direttiva 2005/29/CE sulle pratiche commerciali scorrette, la quale stabilisce che le pratiche commerciali che possono falsare il comportamento economico solo di un gruppo di consumatori vulnerabili in ragione della loro infermità mentale o fisica,

tipi di vulnerabilità siano presi in considerazione nella Proposta (es. vulnerabilità di tipo finanziario). Il problema è di estrema attualità. Negli ultimi anni la letteratura critica ha proposto una nozione più flessibile di vulnerabilità, che non guarda solamente alle caratteristiche degli individui, ma anche a fattori esterni e transitori[26]. Una tale accezione della vulnerabilità sarebbe più che mai giustificata nei contesti di IA basata sui dati (ad es., nel marketing), dove gli individui sono costantemente raggruppati non sulla base di categorie socio-culturali (età, sesso ecc.), bensì sulla base dei loro comportamenti passati e futuri (razionali e non razionali).

Infine, si può sostenere che quando si ha a che fare con sistemi di IA la vulnerabilità è strutturale, e può riguardare potenzialmente chiunque. Questo tipo di vulnerabilità non ha origine nelle caratteristiche personali o circostanziali, ma è inerente al sistema di IA con cui gli individui interagiscono[27]. La prospettiva riguarda il livello di conoscenza tecnica del cittadino medio e l'affidabilità dei sistemi di IA: trattandosi di temi chiave nella strategia europea sull'IA, ci si potrebbe augurare una presa di posizione più ambiziosa sullo sfruttamento delle vulnerabilità.

4.3 Social scoring pubblico

Il terzo divieto riguarda i sistemi di IA finalizzati al cd. *social scoring*, ossia la valutazione e la classificazione dell'affidabilità delle persone fisiche sulla base del loro comportamento in determinati contesti sociali o di altre caratteristiche personali. La norma proibisce la vendita di tali sistemi e il loro utilizzo da parte dell'autorità pubblica (o di altri

della loro età o ingenuità, e tale vulnerabilità è ragionevolmente prevedibile dal professionista, debbano essere valutate nell'ottica del membro medio di tale gruppo.

[26] Si veda ad esempio le teorie proposte nel campo della bioetica da Luna, F., *Elucidating the concept of vulnerability: Layers not labels*, IJFAB: International Journal of Feminist Approaches to Bioethics, 2, 2009, pp. 121–139.; Mackenzie, C., Rogers, W., Dodds, S., *Vulnerability: New essays in ethics and feminist philosophy*, Oxford University Press, 2014.

[27] Burr, C., Cristianini, N., Ladyman, J., *An Analysis of the Interaction Between Intelligent Software Agents and Human Users*, Minds and Machines, 28, 2018, pp. 735-774.

soggetti per conto di essa), quando tale "punteggio sociale" comporti un trattamento pregiudizievole o sfavorevole per individui o gruppi di persone fisiche, in due casi: 1) quando l'effetto dannoso si verifica in contesti sociali diversi da quelli in cui i dati utilizzati dal sistema sono stati originariamente generati o raccolti; oppure 2) quando tale danno è ingiustificato o sproporzionato rispetto al comportamento sociale o alla sua gravità.

La proibizione appare chiaramente ispirata dall'intento di opporsi ai sistemi di *social scoring* diffusi in varia misura nei paesi asiatici tecnologicamente avanzati, ed in particolare in Cina[28]. Il timore della deriva cinese sembra però aver distolto l'attenzione del legislatore dall'utilizzo di sistemi di *scoring* da parte dei privati, che in Europa risulta più frequente. Questa lacuna è altamente problematica considerato che anche lo *scoring* privato può avere un impatto significativo sui diritti fondamentali e i principi democratici di base. Si pensi, ad esempio, all'accesso ai servizi essenziali quali la fornitura di gas e di acqua, ai servizi di accesso al credito o assicurativi, fino alla possibilità di beneficiare di servizi per il tempo libero sulle piattaforme online[29].

Occorre inoltre segnalare la dubbia opportunità dei due requisiti, e cioè quello di integrità contestuale dei dati e della proporzionalità del pregiudizio. Il primo ammette che l'impiego del *social score* sia legittimo se i dati utilizzati dal sistema sono raccolti in un contesto collegato al punteggio sociale. Il secondo, invece, sembra lasciar aperto un ampio margine di discrezionalità nel definire quando il danno è proporzionale rispetto al comportamento sociale.

---

[28] I cittadini cinesi possono per esempio accumulare crediti pagando in tempo le fatture, onorando i contratti stipulati, facendo volontariato nella comunità o donando il sangue, mentre subiscono penalizzazione se non pagano il parcheggio o se non comportano il biglietto nei trasporti pubblici. Il sistema ricompensa i cittadini con vantaggi importanti, quali, ad esempio, l'accesso facilitato a finanziamenti e la facilitazione di viaggi e spostamenti, mentre li penalizza limitando i diritti di voto, escludendoli da scuole private, rallentando la connessione Internet ecc.

[29] Ad esempio, Airbnb ha recentemente brevettato un sistema di IA in grado di generare un punteggio sociale per determinare l'affidabilità dei consumatori sulla base di una serie di dati di social media. Se operativa, questa applicazione di IA porterà probabilmente alla discriminazione di alcuni gruppi sociali nel mercato delle case vacanza, ma non sarebbe oggetto del divieto di cui all'art. 5, lett. c).

4.4 Identificazione biometrica a distanza e in tempo reale

Le ultime pratiche vietate (art. 5(1) lett. d, (2), (3), e (4)) riguardano l'immissione sul mercato, la messa in servizio e l'uso di "sistemi di identificazione biometrica a distanza in tempo reale" in spazi accessibili al pubblico, a scopo di applicazione della legge[30].

Come regola generale, i sistemi di identificazione biometrica a distanza in tempo reale (riconoscimento facciale e vocale) sono vietati. Tuttavia, sono previste molteplici eccezioni, alcune delle quali riguardano specificamente l'uso dell'IA per la giustizia penale, e in particolare per: (i) la ricerca mirata di specifiche vittime potenziali di reato, compresi i bambini scomparsi; (ii) la prevenzione di una minaccia specifica, sostanziale e imminente alla vita o all'incolumità delle persone fisiche o di un attentato terroristico; (iii) l'individuazione, la localizzazione, l'identificazione o il perseguimento dell'autore o del sospetto di una serie di reati specifici[31].

Prevista a fini investigativi, quest'ultima eccezione suscita qualche perplessità: facendo riferimento a una certa gravità del reato (cioè una pena detentiva di almeno tre anni) come stabilita dalla legge di ciascun Stato membro, la disposizione potrebbe infatti portare a disparità di trattamento nei vari Stati membri, con conseguenti complicazioni anche in termini di cooperazione giudiziaria e di polizia[32].

Nei casi menzionati sopra, l'adozione di sistemi di identificazione biometrica a distanza in tempo reale richiede l'adozione di alcune garanzie procedurali. Tali garanzie consistono, da un lato, nella valutazione dei presupposti che autorizzano l'uso di tali tecnologie, e

[30] Tali sistemi sono definiti al punto 37 dell'articolo 3 come sistemi "in cui la cattura dei dati biometrici, il confronto e l'identificazione avvengono senza un ritardo significativo". La definizione è intesa a coprire non solo l'identificazione istantanea, ma anche quella effettuata a breve distanza di tempo; altri tipi di identificazione biometrica sono considerati come "sistema di identificazione biometrica a distanza".

[31] Si fa riferimento ai reati di cui all'articolo 2, paragrafo 2, della decisione quadro 2002/584/GAI del Consiglio e punibile nello Stato membro interessato con una pena o una misura di sicurezza privative della libertà della durata massima di almeno tre anni, come stabilito dal diritto di tale Stato membro.

[32] Lavorgna, A., Suffia, G., *La nuova proposta europea per regolamentare i Sistemi di Intelligenza Artificiale e la sua rilevanza nell'ambito della giustizia penale: un passo necessario, ma non sufficiente, nella giusta direzione*, Diritto Penale Contemporaneo, 2021, pp. 88-103.

delle potenziali conseguenze per i diritti e per le libertà; dall'altro, nelle limitazioni temporali, geografiche e personali da rispettarsi nell'uso di tali sistemi. In ogni caso, si richiede un'autorizzazione preventiva dell'autorità giudiziaria o di un'autorità amministrativa indipendente dello Stato membro e rilasciata su domanda motivata. Tale autorizzazione, tuttavia, "in una situazione di urgenza debitamente giustificata" può essere richiesta anche durante o dopo l'uso. L'autorità competente può concedere l'autorizzazione solo se, sulla base di prove oggettive o di chiare indicazioni presentate, l'uso del sistema in questione è necessario e proporzionato al raggiungimento di uno degli obiettivi consentiti dalla Proposta[33]. Infine, gli Stati membri restano liberi di prevedere la possibilità di autorizzare in via generale l'uso di tali sistemi di identificazione biometrica, sempre nei limiti e alle condizioni di cui ai paragrafi 1, lettera d), 2 e 3.

In forza dell'elevato rischio che questo tipo di sistemi comporta per gli individui e i gruppi, la regolamentazione dettagliata dei "sistemi di identificazione biometrica" è da accogliere con favore. Tuttavia, la portata delle disposizioni appare troppo ristretta e allo stesso tempo doppiamente vaga.

La proibizione dei sistemi biometrici si applica solo all'"identificazione" delle persone fisiche. Il termine non è sempre efficace per descrivere i caratteri di tali tecnologie. Come è noto, molti dei sistemi che utilizzano dati biometrici (ad esempio, il movimento degli occhi e delle labbra, la frequenza cardiaca, la conduttanza cutanea, ecc.) non sono finalizzati all'identificazione dell'individuo (nel senso di identificare la persona fisica con il suo nome e cognome), ma sono utilizzati per il "riconoscimento" degli individui, ossia per categorizzarli secondo classi o metriche predefinite e per valutare il loro comportamento, che può essere considerato pericoloso o indesiderabile. Il problema del perimetro della privacy si ripropone: i sistemi di IA spesso non identificano le persone fisiche, ma sono

[33] Il paragrafo 4 impone agli Stati membri di stabilire nel diritto nazionale norme dettagliate per la richiesta, il rilascio e l'esercizio, nonché il controllo relativo alle autorizzazioni di tali usi.

progettati per "isolarle" dalla massa della popolazione secondo criteri predefiniti[34].

Data la rilevanza delle eccezioni alla regola, infine, il loro perimetro dovrebbe essere oggetto di una ponderata riflessione nel percorso di approvazione del Regolamento. La formulazione attuale è molto ampia, e con ogni probabilità causerà grossi dubbi interpretativi e problemi pratici nei tribunali prima di arrivare ad un accettabile grado di chiarezza.

5 Sistemi IA ad alto rischio: metodologia e obblighi

5.1. Metodologia per definire i sistemi di IA ad alto rischio

I sistemi ad alto rischio sono definiti nella Proposta di Regolamento come sistemi che "creano un rischio elevato per la salute e la sicurezza o i diritti fondamentali delle persone fisiche". L'uso di questi sistemi è permesso, ma regolato da una serie di requisiti e una valutazione di conformità ex ante, di cui diremo nel resto di questa sezione. La definizione di alto rischio è adattata caso per caso, e può comprendere categorie di IA presenti anche nelle pratiche proibite, a seconda dell'uso.

Come per la definizione di IA, la determinazione dell'alto rischio si compone di due parti.

Prima di tutto, un sistema è considerato ad alto rischio se costituisce una componente di sicurezza di un prodotto (o di una sua parte) regolato da normative di armonizzazione dell'Unione Europea (New Legislative Framework, NFL). Le normative di riferimento sono elencate nell'allegato II, e comprendono tra gli altri la direttiva 2009/48/CE sulla sicurezza dei giocattoli e il Regolamento (UE) 2017/745 sui dispositivi medici. Sono inoltre elencate le normative del "vecchio approccio", a cui il Regolamento non si applicherebbe in un

[34] A questo proposito, è interessante notare che questi tipi di sistemi non sono considerati dalla Proposta come ad alto rischio, ma di un livello di rischio inferiore. Essi sono menzionati nell'allegato 3, punto. 1, di cui all'articolo 6, paragrafo 2, come esempio di sistema ad alto rischio (tuttavia, nella descrizione dei sistemi, si fa di nuovo riferimento solo all'identificazione) e nell'articolo 52 che disciplina gli obblighi di trasparenza per alcuni sistemi di IA, e che rappresentano il terzo livello di rischio).

primo momento, ma rispetto al quale i requisiti per i sistemi di IA ad alto rischio dovranno essere presi in considerazione al momento dell'adozione di nuove norme in materia. Di questa ultima categoria fanno parte il Regolamento (UE) 2018/858 sull'aviazione civile e il Regolamento (UE) n. 168/2013 sui veicoli a motore.

La seconda parte della definizione dell'alto rischio è costituita anche in questo caso da un elenco. L'elenco è contenuto nell'Allegato III e include otto macro-aree in cui i sistemi di IA utilizzati si presumono di alto rischio. Tra le macro-aree sono incluse l'identificazione biometrica, l'accesso a servizi pubblici e privati, attività di contrasto, e amministrazione della giustizia[35].

I sistemi che rientrano in queste categorie sono soggetti a una serie rigorosa di regole valide sia per il fornitore che per l'utilizzatore, che ricomprendono: la creazione di un sistema di addestramento e validazione (art. 9), un obbligo di qualità e accuratezza dei dati (art. 10), documentazione tecnica, e requisiti di trasparenza verso l'utilizzatore (art 13)[36].

Di particolare interesse nell'ambito della giustizia penale sono i sistemi descritti nella categoria *attività di contrasto* (numero 6, lettera d dell'allegato III): "sistemi di IA destinati a essere utilizzati dalle autorità di contrasto [...] per la valutazione dell'affidabilità degli elementi probatori nel corso delle indagini o del perseguimento di reati", "per individuare i deep-fake", "per la profilazione delle persone fisiche", e "per l'analisi criminale riguardo alle persone fisiche, che consentono alle autorità di contrasto di eseguire ricerche in set di dati complessi, correlati e non correlati, resi disponibili da fonti di dati diverse o in formati diversi, al fine di individuare modelli sconosciuti o scoprire relazioni nascoste nei dati".

Anche per quanto riguarda la determinazione dei sistemi ad alto rischio, e specificamente alle categorie contenute nell'Allegato II, la Proposta di Regolamento non sembra seguire propriamente un

[35] Le altre aree sono: la gestione e funzionamento delle infrastrutture, l'istruzione e formazione professionale, l'occupazione, gestione dei lavoratori e accesso al lavoro autonomo, la gestione della migrazione, dell'asilo e del controllo delle frontiere.

[36] Bisogna chiarire che quando il Regolamento parla di utilizzatore non intende il soggetto ultimo, ma un intermediario, che offre servizi che fanno uso di sistemi di IA, anche laddove il sistema di IA sia stato sviluppato da terzi.

approccio *risk-based*. La Proposta seleziona determinate aree di mercato nelle quali l'introduzione di sistemi di IA implica presumibilmente un alto livello di rischio per i diritti e le libertà fondamentali degli individui. Le aree connesse al rischio sono certamente sensibili: tuttavia, non si individua un metodo chiaro e generale per determinare le specifiche caratteristiche che darebbero luogo al rischio di un sistema. Perciò, anche quando un sistema di IA adottato negli ambiti summenzionati non genera necessariamente un rischio (si pensi al motore di ricerca giuridico utilizzato dal giudice), questo potrebbe essere considerato ad alto rischio.

In aderenza all'approccio *risk-based*, ci si sarebbe aspettati una previsione più simile all'istituto della valutazione di impatto sulla protezione dei dati contenuta nel GDPR. Questa costituisce una procedura generale di valutazione e mitigazione dell'impatto del trattamento sui dati personali, nel momento in cui tale trattamento comporta un rischio elevato per i diritti e le libertà delle persone interessate. I criteri di rischio sono stati individuati attraverso le linee guida elaborate dal Gruppo Art. 29 ed includono trattamenti valutativi o di *scoring*, monitoraggio sistematico, combinazione o raffronto di insiemi di dati derivanti da più trattamenti con finalità diverse, dati relativi a soggetti vulnerabili, utilizzo di tecnologie innovative come riconoscimento facciale, e così via[37].

5.2 Governance dei dati

Dopo aver definito le regole di classificazione per i sistemi di IA ad alto rischio, la Proposta di Regolamento contiene una lunga lista di requisiti essenziali (Capo 2) che si ricollega agli obblighi dei soggetti previsti nel capo successivo. La maggioranza degli obblighi ricade sul "fornitore", ovvero la persona o ente che sviluppa un sistema IA al fine

[37] Gruppo Art. 29, *Guidelines on Data Protection Impact Assessment (DPI A) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Adopted on 4 April 2017 (As last Revised and Adopted on 4 October 2017)*, European Commission, 2017. Per altro si potrebbe sostenere che la DPIA prevista dal GDPR e come interpretata dal Gruppo dell'Art. 29 sarebbe già sufficiente a includere i rischi derivanti dall'utilizzo di sistemi decisionali automatizzati, quindi anche legati a sistemi di IA.

di immetterlo sul mercato o di metterlo in servizio con il proprio nome o marchio, a titolo oneroso o gratuito (art. 3, n. 3).

In via generale, i fornitori di sistemi di IA ad alto rischio devono costituire un meccanismo di gestione della qualità che garantisca conformità ad una serie di requisiti. Tra questi requisiti figurano: un'appropriata governance dei dati (art. 10); la conservazione delle registrazioni (art. 12); la trasparenza e fornitura di informazioni agli utenti (art. 13).

Secondo l'art. 10, i sistemi di IA ad alto rischio basati su tecniche di apprendimento dai dati devono soddisfare alcuni criteri di qualità, in particolare per quanto riguarda i dati per l'addestramento (*data set*), per la convalida (*validation set*) e per la prova (*test set*). I dati devono essere pertinenti e rappresentativi, e devono tenere conto delle proprietà specifiche dell'area geografica di applicazione. In modo piuttosto ambizioso, il legislatore UE richiede inoltre che i dati siano "esenti da errore" e "completi". Occorre notare come un tale livello di qualità potrebbe essere difficile da richiedere a livello tecnico. Non si rintraccia invece (ma risulterebbe utile) la previsione di misure idonee a garantire la minimizzazione degli errori.

Un'ulteriore pratica adeguata di governance dei dati è costituita dal monitoraggio, il rilevamento e la correzione delle distorsioni (*bias*). È ormai noto, infatti, come l'utilizzo di sistemi di IA basati sull'apprendimento automatico aumenti il rischio di casi di discriminazione indiretta, anche non intenzionale. Ciò può accadere i molti modi, a seconda di come vengano definite le variabili target del modello, di come vengano selezionate le caratteristiche del modello di apprendimento, e degli elementi sostitutivi (*proxy*)[38]. Al fine di minimizzare il rischio di esiti discriminatori, la Proposta prevede che un eventuale trattamento di categorie speciali di dati di cui all'art. 9 GDPR possa avvenire in modo legittimo, fatte salve le tutele necessarie per i diritti e le libertà fondamentali delle persone fisiche[39]. Spesso,

[38] Barocas, A. D. Selbst, *Big data's disparate impact*, Calif. L. Rev., 104, 2016, p. 671.

[39] Per garantire armonia tra le due discipline, occorrerebbe però modificare anche l'art. 9 GDPR, che allo stato attuale non stabilisce come requisito di legittimità del trattamento di dati sensibili il loro uso per il rilevamento di pregiudizi. Allo stesso tempo però occorrerebbe definire con maggior precisione nella Proposta di

infatti, è difficile rilevare il potenziale di distorsioni nei sistemi di apprendimento automatico senza conoscere le caratteristiche protette degli individui (i fattori di rischio della discriminazione), che spesso corrispondono a dati sensibili[40]. In ogni caso, l'esenzione si applica solamente ai soggetti fornitori di sistemi di IA ad alto rischio e dunque non può fornire un *passe-partout* per eventuali soggetti che, a monte, collezionano dati personali per lo sviluppo di tali sistemi.

Infine, l'ultimo comma dell'art. 10 richiede l'implementazione di pratiche adeguate di governance dei dati anche per i sistemi di IA ad alto rischio che non si basano sull'addestramento automatico. Quest'ultimo aspetto è apprezzabile, considerato che anche altre tecniche di IA, come ad esempio sistemi esperti o alcuni modelli di ottimizzazione, possono condurre a effetti pregiudizievoli per i diritti e le libertà fondamentali.

5.3 Trasparenza e spiegabilità

I sistemi di IA ad alto rischio devono essere progettati in maniera da rendere comprensibile il loro funzionamento a utenti[41] e fornitori. Questi devono avere chiare istruzioni relative alle finalità del sistema e alla supervisione umana[42].

Occorre evidenziare come l'uso che viene fatto del termine *trasparenza* in questa sezione della Proposta è riferito alla necessità di chiarezza verso l'utente professionista che deve poter comprendere il funzionamento del sistema e il suo risultato. Questa esigenza di

Regolamento le procedure di monitoraggio, rilevamento e correzione delle distorsioni, al fine di non lasciare aperta la porta a forme di trattamento non indispensabili di categorie di dati speciali.

[40] I. N. Cofone, *Algorithmic discrimination is an information problem*, Hastings LJ, 70, 2018, p. 1389.

[41] "Qualsiasi persona fisica o giuridica, autorità pubblica, agenzia o altro organismo che utilizza un sistema di IA sotto la sua autorità". Questa definizione si discosta da quella usata nel linguaggio comune: in questo caso l'utente non comprende l'utilizzatore finale, ossia colui che interagisce con il sistema predisposto da altri, senza avere autorità o controllo su di esso.

[42] R. Chatila, V. Dignum, M. Fisher, F. Giannotti, K. Morik, S. Russell, K. Yeung, *Trustworthy AI*, in *Reflections on Artificial Intelligence for Humanity*, Springer, 2021, pp. 13–39

trasparenza differisce dal tipo di trasparenza previsto nel Titolo IV, di cui si parlerà nella sezione 6.

Si può dubitare che l'obbligo di trasparenza previsto per i sistemi di alto rischio, richiedendo l'accesso e la comprensione di informazioni tecniche, possa sempre ed in ogni caso diminuire il rischio per gli utenti. Spesso i soggetti nella catena di utilizzazione dei prodotti o servizi di IA non hanno conoscenze tecniche che permettano la comprensione di queste informazioni: si pensi al business che utilizza i sistemi di *inferential analytics* di Google per la pubblicità mirata. Inoltre, la vaghezza nella descrizione dei requisiti lascia aperte alcune questioni. L'art. 13 si concentra sull'interpretabilità dell'output dei sistemi di IA da parte dei suoi utenti senza chiarire il concetto di interpretabilità e la sua correlazione con la spiegabilità[43]. Inoltre, non si rinvengono riferimenti all'utilizzo di tecniche di *explainable AI*[44].

Al Considerando 70 della Proposta, relativamente ai sistemi ad alto rischio e a quelli che più in generale comportano interazioni con persone fisiche, si richiede di assicurare trasparenza nei confronti del pubblico. Tuttavia, nel testo del Regolamento non si ha riscontro di un obbligo generale di trasparenza/spiegabilità nei confronti delle persone fisiche utilizzatori finali dei sistemi. Nelle linee guida etiche del Gruppo di esperti sull'Intelligenza artificiale [45] si parlava espressamente del pubblico, e dei requisiti di trasparenza e spiegabilità [46]. La trasparenza e la spiegabilità del sistema sono elementi fondamentali per i cittadini consumatori perché essi possano capire la motivazione di una decisione algoritmica[47], e quindi accettarla

[43] G. Sartor, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, Study PE 641.530, European Parliamentary Research Service, 2020.

[44] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, *A survey of methods for explaining black box models*, ACM computing surveys (CSUR), 51, 2018, pp. 1–42.

[45] Gruppo di esperti ad alto livello sull'Intelligenza artificiale, Orientamenti etici, cit..

[46] Affinché un sistema di IA possa essere tecnicamente spiegabile gli esseri umani devono poter capire e tenere traccia delle decisioni prese dal sistema stesso

[47] M. Brkan, *Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond*, International journal of law and information technology, 27, 2019, pp. 91–121.

o eventualmente contestarla. Questo approccio è stato perso nella Proposta di Regolamento, e si è prediletto un approccio incentrato sulla trasparenza verso i produttori e fornitori. Si sarebbe potuto specificare l'obbligo di trasparenza con riguardo alla logica coinvolta nella decisione algoritmica di cui all'art. 13 del GDPR, oppure si sarebbe potuto sancire ufficialmente il tanto discusso diritto ad ottenere una spiegazione [48] . Nessuna di queste soluzioni è stata presa in considerazione, e un ripensamento in questo senso sarebbe forse opportuno.

5.4 Supervisione umana

All'art. 14, la Proposta stabilisce che i sistemi ad alto rischio siano sviluppati "anche con strumenti di interfaccia uomo-macchina adeguati, in modo tale da poter essere efficacemente supervisionati da persone fisiche".

La sorveglianza, che mira a prevenire o minimizzare i rischi per la salute, la sicurezza, o i diritti fondamentali (art 14 (2)), deve essere garantita mediante misure individuate dal fornitore, e integrate nel sistema dal fornitore stesso o implementate dall'utente. Tali misure mirano a supportare le azioni indicate al paragrafo 4 dell'art. 14: a) comprendere appieno le capacità e i limiti del sistema di IA ed essere in grado di monitorarne il funzionamento; b) mantenere un atteggiamento consapevole rispetto al rischio di *bias*; c) essere in grado di interpretare correttamente l'output del sistema; d) essere in grado di decidere di non usare il sistema o di ignorare, annullare, o ribaltare il suo output; e) essere in grado di intervenire sul funzionamento del sistema o di interromperlo mediante un pulsante di "arresto" o una procedura analoga. Inoltre, l'art. 29 della Proposta, tra gli obblighi degli utenti dei sistemi di IA ad alto rischio, elenca l'obbligo di usare il sistema (art. 29 (1)) e monitorarne il funzionamento (art. 29 (3)) sulla base delle istruzioni per l'uso.

[48] S. Wachter, B. Mittelstadt, L. Floridi, *Why a right to explanation of automated decision-making does not exist in the general data protection regulation*, International Data Privacy Law, 7, 2017, pp. 76–99; G. Malgieri, G. Comandé, *Why a right to legibility of automated decision-making exists in the general data protection regulation*, International Data Privacy Law, 2017.

Nel loro insieme, queste norme appaiono in linea con i documenti precedenti prodotti a livello europeo, come ad esempio gli Orientamenti etici per un'Intelligenza artificiale affidabile, che promuovono il requisito della sorveglianza umana[49]. Tuttavia, risulta difficile comprendere come l'individuo posto a sorveglianza del sistema possa, per tutti i tipi e le applicazioni di IA, esercitare effettivamente le azioni descritte nel paragrafo 4 dell'art. 14, sulla base di una perfetta comprensione delle capacità, limiti, e output del sistema, per di più operando una sintesi tra esigenze di tutela tanto ampie e diverse quali sicurezza, salute, diritti umani (che potrebbero richiedere anche una difficile operazione di bilanciamento). In realtà, le possibilità per l'umano di comprendere appieno le capacità e i limiti del sistema è ad oggi assai limitata per molte tipologie di IA, e porterebbe di conseguenza, secondo alcuni esperti, a un divieto indiretto di tali sistemi[50].

Come sottolineato da Selbst, l'Intelligenza artificiale (in particolare quella basata su sistemi di apprendimento automatico) è fondamentalmente aliena per un essere umano. Di norma, anzi, lo stesso scopo di tali sistemi è quello di imparare a compiere operazioni secondo modalità in larga parte precluse agli esseri umani[51]. Infatti, i sistemi di IA sono di solito sviluppati e introdotti sul presupposto che, per i compiti che sono loro assegnati, sorpassino le capacità degli umani, eccedendo non solo i loro limiti cognitivi, ma anche quelli temporali nell'accedere alle informazioni, elaborarle e prendere decisioni. Questo è tanto più vero se consideriamo che, nel campo delle valutazioni statistiche, la capacità di giudizio umana è inferiore a quella dei modelli matematici[52].

---

[49] Gruppo di esperti ad alto livello sull'Intelligenza artificiale, Orientamenti etici, cit. p. 17.

[50] M. Ebers, V. R. Hoch, F. Rosenkranz, H. Ruschemeier, B. Steinrötter, *The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)*, J, 4, 2021, pp. 589-603.

[51] A. D. Selbst, *Negligence and AI's human users*, BUL Rev., 100, 2020, p. 1315.

[52] C. Deskus, *Fifth Amendment Limitations on Criminal Algorithmic Decision-Making*, NYUJ Legis. & Pub. Policy, 21, 2018, p. 237.

Il problema di fondo, quindi, è uno sfasamento tra competenze e responsabilità: i sistemi di IA sono adottati per le loro capacità di decisione e previsione superiori rispetto agli esseri umani: quegli stessi esseri umani, tuttavia, sono poi incaricati di giudicare la qualità dei sistemi stessi, creando quello che spesso è un "compito impossibile"[53].

Anche quando l'uomo mantiene formalmente il controllo sulla decisione finale, la possibilità che egli sia in grado di entrare efficacemente nel merito della decisione rischia di rimanere un'ipotesi remota. Più verosimilmente, il supervisore umano tenderà a fare affidamento sulle decisioni di un sistema di IA[54], tanto più se questo è stato certificato, a meno che non abbia motivi specifici per ritenere che esso sia malfunzionante, o che non sia in grado di incorporare nella sua valutazione ulteriori elementi, esterni al sistema stesso[55].

6 Obblighi di trasparenza per determinati sistemi: i cd. "sistemi di medio rischio"

Il titolo IV comprende il solo articolo 52 e affronta tre specifici sistemi di IA che, nell'approccio basato sul rischio, richiedono l'introduzione di obblighi di trasparenza. Questi sistemi sono visti come problematici per il loro potenziale manipolativo, e per questo motivo è previsto un obbligo di trasparenza nei confronti della persona fisica che interagisce col sistema.

In proposito, occorre effettuare qualche precisazione. Innanzitutto, la trasparenza richiesta per questi tipi di IA non coincide con la trasparenza intesa come "spiegabilità" richiesta dall'art. 13 per i sistemi ad alto rischio, di cui si è detto più sopra (sezione 5.3). Ciò che si richiede qui è di rendere noto a chi lo utilizza che si sta interagendo con un sistema di IA. Difatti, i beneficiari dell'art. 52 non sono gli "utenti" nel senso utilizzato dalla Proposta, ma tutte le "persone

[53] L. Bainbridge, *Ironies of automation, in Analysis, design and evaluation of man–machine systems*, Elsevier, 1983, pp. 129–135.

[54] J. Millar, I. Kerr, *Delegation, relinquishment, and responsibility. The prospect of expert robots*, in Robot Law, Edward Elgar Publishing, 2016

[55] F. Lagioia, G. Contissa, *The Strange Case of Dr. Watson: Liability Implications of AI Evidence-Based Decision Support Systems in Health Care*, Eur. J. Legal Stud., 12, 2020, p. 281.

fisiche" che con il sistema interagiscono, e specialmente gli "utilizzatori finali". Infine, quando i sistemi soggetti all'art. 52 rientrino nella categoria ad alto rischio, a tali obblighi si aggiungono i requisiti previsti dal titolo III (inclusa la spiegabilità).

Il primo gruppo descritto nell'art. 52 ricomprende i sistemi "destinati a interagire con le persone fisiche". Sebbene la definizione appaia piuttosto ambigua, la disposizione è stata pensata per affrontare il problema dei chatbot e altri sistemi di IA personificati[56]. In questo caso, il fornitore del sistema deve garantire che le persone fisiche che interagiscono con il sistema (gli "utilizzatori finali") siano consapevoli di interagire con un sistema di IA. L'obbligo non si applica quando l'interazione "è evidente dalle circostanze e dal contesto di utilizzo"[57]. Rimangono inoltre esclusi dall'obbligo i sistemi messi a disposizione al pubblico per segnalare un reato.

Il secondo gruppo ricomprende i sistemi di riconoscimento delle emozioni e di categorizzazione biometrica. Questi sistemi interpretano i dati biometrici e fanno valutazioni sugli individui ma, a differenza dei sistemi di cui all'art. 5 lett. d) e all'Allegato III, n. 1, non sono finalizzati all'identificazione della persona fisica. Alla prima categoria possono essere ricondotti, ad esempio, i sistemi di IA che applicano tecniche di analisi del sentimento o di profilazione psicografica. Alla seconda categoria possono essere ricollegati invece alcuni *wearable* come Apple SmartWatch o FitBit, che utilizzano dati biometrici per effettuare valutazioni sullo stato di salute degli utenti, ma anche sistemi di riconoscimento vocale (come Siri o Alexa) o facciale. Per quanto certamente utile, si può dubitare che siffatto requisito di trasparenza possa diminuire in modo significativo la pericolosità di questi sistemi rispetto a diritti e libertà fondamentali.

---

[56] I chatbot possono essere definiti come sistemi software che utilizzano tecniche di elaborazione del linguaggio naturale (NLP) e di generazione (NLG) per simulare una conversazione con un essere umano e interagire con gli utenti attraverso interfacce basate sul dialogo.

[57] Ragionevolmente, questa disposizione esclude dal campo di applicazione le interazioni con i robot (come Alexa o RobotClean) dove la natura artificiale del sistema è visibile apertamente. Diversi sono i casi in cui la natura artificiale del sistema potrebbe essere concepita dal fornitore, come nell'ambito dell'e-commerce o del marketing: il consumatore potrebbe pensare di stare interagendo con un venditore umano mentre invece sta interagendo con un sistema IA.

Infine, il terzo gruppo di sistemi inclusi nell'articolo 52 si riferisce alla pratica dei "deep fakes". Tale pratica è definita come l'uso del sistema di IA per generare o manipolare un contenuto di immagine, audio o video che assomiglia sensibilmente a persone, luoghi, oggetti o altre entità o eventi esistenti e che apparirebbe falsamente a una persona come autentico o veritiero.

Per tutte le tipologie di sistemi qui descritte è prevista un'eccezione all'obbligo di trasparenza nel caso in cui i sistemi siano utilizzati a fini di indagine e accertamento di reati.

7. Valutazione di conformità, monitoraggio e governance del sistema

Per completare il quadro della Proposta, occorre infine soffermarci sulle forme di controllo preventivo dei sistemi di IA e di monitoraggio successivo alla loro immissione nel mercato, nel quadro più generale dei sistemi di governance previsti dal legislatore europeo.

Le nuove norme prevedono che, prima di poter entrare sul mercato, i sistemi di IA ad alto rischio siano sottoposti ad una preliminare valutazione di conformità ai requisiti del Regolamento. Come regola generale, l'art. 43(2) stabilisce che valutazione di conformità si basa sul controllo interno del fornitore (disciplinato dall'Allegato VI). Ciò significa che i fornitori valutano autonomamente l'aderenza del loro sistema di gestione della qualità e della documentazione tecnica specifica ai requisiti normativi e agli standard comuni. A questo fine, i fornitori possono adeguarsi agli standard esistenti[58] oppure possono implementare a livello tecnico i requisiti essenziali del Regolamento in modo autonomo.

In determinati casi è invece previsto l'intervento di soggetti terzi. Per i sistemi di categorizzazione biometrica delle persone fisiche (art. 43(1)), l'autovalutazione del fornitore richiede l'approvazione di un "organismo notificato" (art. 33) indipendente e accreditato presso le autorità di sorveglianza prevista dal Regolamento. I sistemi che costituiscono componenti di sicurezza di prodotti già regolati dal New

[58] Su delega della Commissione europea, gli organismi di standardizzazione (come il CEN o il CENELEC) possono prevedere degli standard tecnici relativi ai requisiti essenziali dei sistemi di IA ad alto rischio.

Legislative Framework (NLF) seguono la procedura di valutazione di cui alla relativa normativa, al fine di evitare duplicazioni procedurali (art. 43(3)).

Al di là di questi casi, dunque, per molti dei sistemi di IA ad alto rischio i fornitori potranno apporre il marchio di conformità UE sulla base della sola autovalutazione. Questa scelta solleva non poche perplessità, nella misura in cui si esclude qualsiasi supervisione preventiva relativamente a tecnologie considerate molto rischiose per le persone[59].

La Proposta di Regolamento stabilisce un sistema di governance dell'ordinamento sia a livello UE che a livello nazionale. A livello dell'UE, la Proposta istituisce una nuova autorità, il Comitato europeo per l'Intelligenza artificiale, con compiti di coordinamento della governance del sistema (titolo VI, capo 1). Il Comitato garantisce l'efficace cooperazione tra le autorità nazionali (anche attraverso la condivisione di *best practices*), e sostiene l'attività della Commissione formulando pareri e linee guida.

Gli Stati membri, dal canto loro, dovranno designare una o più autorità nazionali competenti, per sorvegliare sull'applicazione e il rispetto del Regolamento (capo 2). Le autorità potranno chiedere ai fornitori informazioni e accedere alla documentazione, sollecitare l'adozione di misure correttive e il ritiro del sistema dal mercato.

Per quanto riguarda gli ulteriori strumenti di governo del sistema, il titolo VII della Proposta dispone l'istituzione, presso la Commissione europea, di una banca dati per i sistemi di IA indipendenti ad alto rischio. La banca dati raccoglierà una serie di informazioni importanti sulle caratteristiche e le finalità dei sistemi, sulle certificazioni pertinenti e sull'identità del fornitore[60]. Si prevede poi, al titolo IX, che la Commissione e gli Stati membri promuovano l'adozione di codici di condotta per i sistemi di IA diversi dai sistemi ad alto rischio. Tali codici dovrebbero adeguarsi ai requisiti previsti per i sistemi ad alto rischio, in modo che anche i sistemi non ad alto rischio rispettino gli stessi standard di quelli ad alto rischio.

---

[59] Per un'analisi della valutazione di conformità v. J. MÖKANDER, M. AXENTE, F. CASOLARI, and L. FLORIDI, *Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI Regulation*, Minds and Machines, 2021, pp. 1–28.

[60] L'elenco completo delle informazioni è contenuto nell'Allegato VII.

Per quanto riguarda le sanzioni, la Proposta sembra seguire l'approccio già adottato dal GDPR. Al titolo X sono previste sanzioni amministrative in caso di violazione o mancato rispetto delle norme contenute nel Regolamento. In caso di violazione dei divieti di cui all'articolo 5 o della non conformità del sistema di IA ai requisiti dell'articolo 10, le sanzioni possono arrivare fino a trenta milioni di euro o, se l'autore del reato è una società, fino al 6% del fatturato annuo globale. In caso di altre violazioni, le sanzioni sono ridotte ad un massimo di venti milioni di euro o del 4% del fatturato della società (art. 71).

Nonostante la rilevanza delle misure sanzionatorie e la possibilità per le autorità di richiedere il ritiro dal mercato o il richiamo dei prodotti, la disciplina complessiva non appare del tutto sufficiente a garantire un'applicazione efficace. L'esperienza del GDPR mostra che un eccessivo affidamento sull'*enforcement* delle autorità nazionali si espone al rischio di difformità applicative, a causa delle notevoli divergenze di risorse e prassi di intervento nei diversi Stati membri. Com'è usuale in campo tecnologico, poi, si pone il problema di possibili difformità nei criteri di valutazione degli standard tecnici, che sarà tanto più pregnante in un settore, quello dell'IA, caratterizzato da un elevato numero di attori e di standard diversi.

Infine, occorre segnalare che la Proposta non contiene meccanismi di azione per gli individui o gruppi danneggiati, nemmeno nella forma di reclamo (ad es. all'autorità di sorveglianza del mercato. L'assenza di tali previsioni, che avrebbero compensato i limiti dell'*enforcement* pubblico, risulta certamente criticabile, avendo come effetto anche il ridimensionamento del ruolo della società civile nello sviluppo affidabile dell'IA.

8. Conclusioni

Sulla base di questa breve rassegna, possiamo formulare qualche osservazione conclusiva. La Proposta costituisce, come si è visto, il frutto di un processo condiviso nell'alveo delle istituzioni europee, ed ha il merito di riunire i diversi aspetti tecnologici, socio-economici e regolativi sottesi allo sviluppo e alla diffusione dell'IA in Europa. Elabora una definizione e una classificazione dei diversi sistemi, fornisce un insieme di requisiti e di incombenze per quei sistemi che

mettono a rischio i diritti e i valori europei, stabilisce un'architettura di governance e monitoraggio delle tecnologie. Tuttavia, diversi sono i punti deboli di questa legislazione.

Si sono notate le carenze relative ai diversi aspetti della normativa: le incertezze nella definizione e classificazione dei sistemi (si pensi alla genericità del requisito dei "metodi statistici), le ambiguità nelle proibizioni (come il *social scoring* pubblico) e nei requisiti dei sistemi ad alto rischio (ad es. la sorveglianza umana), le insufficienze nei meccanismi di controllo (si pensi al cd. controllo interno). A nostro parere, tutti questi aspetti sottendono due elementi di carattere generale che rivelano la prospettiva complessiva del legislatore europeo, da un lato volta alla protezione dei diritti e delle libertà fondamentali, ma dall'altro largamente influenzata dal modello di legislazione industriale orientata al mercato.

Il primo elemento è costituito dall'approccio *top-down* del Regolamento [61] . Il Legislatore identifica le tecnologie dell'AI, predetermina le pratiche proibite e rischiose e identifica ex ante il livello di rischio, facendone derivare le regole applicabili. Tuttavia, lo studio dell'automazione suggerisce come sia più adeguato, per la regolazione di tali tecnologie, un approccio *bottom-up*, basato sull'identificazione delle caratteristiche della tecnologia e dei processi nei quali essa viene utilizzata (e ciò soprattutto se si opera in un contesto basato sul rischio). In caso contrario, come si è visto nel presente commento, si corre il pericolo di creare categorie normative astratte che non rappresentano in modo adeguato caratteri e rischi effettivi delle singole soluzioni tecnologiche.

In secondo luogo, l'impianto complessivo della Proposta rivela una contraddizione di fondo fra il piano pubblicistico e quello dei diritti individuali. Le norme intendono regolare il rischio che l'IA crea per i valori e i diritti umani tutelati dall'Unione. E tuttavia: dove si trova l'essere umano in queste norme? La Proposta è incentrata su obblighi e prerogative del fornitore, dell'utente e del supervisore: definisce i requisiti di sviluppo dei sistemi, le modalità di utilizzo e certificazione, i meccanismi di controllo. Dei soggetti su cui ricadono le conseguenze dell'utilizzo del sistema non v'è quasi traccia: a chi subisce l'operato

[61] L. FLORIDI, *The European Legislation on AI: a Brief Analysis of its Philosophical Approach*, Philosophy & Technology, 34, 2021, pp. 215–222.

del sistema (ad esempio chi si ritrova sorvegliato, indagato, accusato da un'IA) non si conferiscono diritti né strumenti di azione. Il Regolamento proclama di volere difendere i valori europei e i diritti fondamentali dai rischi dell'IA: tuttavia, per perseguire questo lodevole obiettivo, nessuno strumento specifico di difesa in più viene fornito ai titolari dei diritti coinvolti.

Perché non mettere nero su bianco nella Proposta, ad esempio, il diritto ad ottenere una spiegazione da parte del sistema AI? O il diritto a non essere soggetti a decisioni discriminatorie? Per entrambi gli aspetti, il legislatore avrebbe potuto fornire linee guida di sviluppo tecnico e certificazione. Si tratta di due aspetti, peraltro su cui il GDPR non aveva fatto chiarezza, e questa poteva certamente essere l'occasione per rimediare[62].

L'attenzione all'impatto dell'IA sulle persone, del resto, sarebbe forse stata agevolata da un approccio *bottom-up*. Si sarebbe potuto, ad esempio, partire dalla Carta dei diritti fondamentali dell'Unione europea, stabilendo per ciascun diritto i parametri di misura del rischio legato alla sua violazione (tenendo conto del "peso" di ogni diritto, e del grado di rischio). Sulla base di questi parametri, si sarebbe potuto determinare, di volta in volta, il livello di rischio dello specifico sistema di IA (evitando così il "*cherry-picking*" di cui si è detto), nonché le misure di protezione individuali. In questo modo, sarebbe stato possibile fornire una sorta di diritti "accessori" che rendessero "azionabili" i diritti umani messi in pericolo.

Si può insomma concludere che Proposta di Regolamento, tentando di coniugare i diritti individuali con le esigenze di mercato relative all'IA, fatichi a trovare un bilanciamento soddisfacente e soluzioni applicative adeguate. Ci si può certamente augurare che il prosieguo del percorso legislativo conduca ad un maggiore equilibrio.

[62] S. WACHTER, B. MITTELSTADT, L. FLORIDI, *op. cit.*; M. BRKAN, *op. cit.*, pp. 100-101.

Bibliografia

Bainbridge, L., Ironies of automation, in Analysis, design and evaluation of man–machine systems, Elsevier, 1983, pp. 129–135.

Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev., 104, 2016, p. 671.

BEUC, EU Consumer protection 2.0. Structural asymmetries in digital consumer market, Report, March 2021.

Bibal, M. Lognoul, A. De Streel, B. Frénay, Legal requirements on explainability in machine learning, Artificial Intelligence and Law, 29 (2021), pp. 149–169.

Brkan, M. Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond, International Journal of Law and Information Technology, 27, 2019, pp. 91–121.

Burr, N. Cristianini, Can Machines Read our Minds?, Minds and Machines, (2019), pp. 1-34.

Burr, N. Cristianini, J. Ladyman, An Analysis of the Interaction Between Intelligent Software Agents and Human Users, Minds and Machines, 28, 2018, pp. 735-774.

Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., Yeung, K., Trustworthy AI, in Reflections on Artificial Intelligence for Humanity, Springer, 2021, pp. 13–39

Cofone, I., Algorithmic discrimination is an information problem, Hastings LJ, 70, 2018, p. 1389.

Deskus, C., Fifth Amendment Limitations on Criminal Algorithmic Decision-Making, NYUJ Legis. & Pub. Policy, 21, 2018, p. 237.

Dignum, V., Responsible artificial intelligence: how to develop and use AI in a responsible way, Springer Nature, 2019.

Ebers, M., Hoch, V. R., Rosenkranz, F., Ruschemeier, H., Steinrötter, B., The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS), J, 4, 2021, pp. 589-603.

Floridi, L., The European Legislation on AI: a Brief Analysis of its Philosophical Approach, Philosophy & Technology, 34, 2021, pp. 215–222.

Gabriele M., (DG CONNECT), 'A European Strategy for Artificial Intelligence' (2nd ELLIS Workshop in Human-Centric Machine Learning (YouTube recording), 10 Maggio 2021) https://youtu.be/OZtuVKW qhl0?t=10346.

Galli, F., Godano, F., Il rapporto di lavoro dei riders e la natura discriminatoria delle condizioni di accesso al lavoro dell'algoritmo frank, Diritto di Internet, III, 2021, pp. 275–288.

Glauner, P., An Assessment of the AI Regulation Proposed by the European Commission, 2021.

Gruppo Art. 29, Guidelines on Data Protection Impact Assessment (DPIA), Adopted on 4 April 2017 (As last Revised and Adopted on 4 October 2017), European Commission, 2017.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti G., Pedreschi, D., A survey of methods for explaining black box models, ACM computing surveys (CSUR), 51, 2018, pp. 1–42.

Lagioia, F., Contissa, G. The Strange Case of Dr. Watson: Liability Implications of AI Evidence-Based Decision Support Systems in Health Care, Eur. J. Legal Stud., 12, 2020, p. 281.

Lavorgna, G. Suffia, La nuova proposta europea per regolamentare i Sistemi di Intelligenza Artificiale e la sua rilevanza nell'ambito della giustizia penale: un passo necessario, ma non sufficiente, nella giusta direzione, Diritto Penale Contemporaneo, 2021, pp. 88-103.

Luguri, L. Strahilevitz, Shining a light on dark patterns, University of Chicago, Public Law Working Paper, 2019.

Luna, F., Elucidating the concept of vulnerability: Layers not labels, IJFAB: International Journal of Feminist Approaches to Bioethics, 2, 2009, pp. 121–139.;

Mackenzie, W. Rogers, S. Dodds, Vulnerability: New essays in ethics and feminist philosophy, Oxford University Press, 2014.

Malgieri, G., Comandé, G., Why a right to legibility of automated decision-making exists in the general data protection regulation, International Data Privacy Law, 2017.

Matz, S. C., Kosinski, M., Nave, G., Stillwell, D. J., Psychological targeting as an effective approach to digital mass persuasion, Proceedings of the National Academy of sciences, 114, 2017, pp. 12714-12719.

Millar, J., Kerr, I., Delegation, relinquishment, and responsibility. The prospect of expert robots, in Robot Law, Edward Elgar Publishing, 2016

Packard, V., Payne, R. The hidden persuaders, McKay New York, 1957.

Sartor, G., The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, Study PE 641.530, European Parliamentary Research Service, 2020.

Selbst, D., Negligence and AI's human users, BUL Rev., 100, 2020, p. 1315.

Wachter, S., Mittelstadt, B., Floridi, L., Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law, 7, 2017, pp. 76–99;

Yeung, K., 'Hypernudge': Big Data as a mode of regulation by design, Information, Communication &; Society, 20, 2017, pp. 118-136.