ISSN 1826-3534

federalismi.it
RIVISTA DI DIRITTO PUBBLICO ITALIANO, COMPARATO, EUROPEO

30 GIUGNO 2021

# La black box: l'esplicabilità delle scelte algoritmiche quale garanzia di buona amministrazione

di Germana Lo Sapio
Magistrato T.A.R. Campania

# La black box: l'esplicabilità delle scelte algoritmiche quale garanzia di buona amministrazione*

**di Germana Lo Sapio**
Magistrato T.A.R. Campania

**Abstract [It]:** L'esame della questione giuridica della esplicabilità delle decisioni amministrative, che si avvalgono di modelli di Intelligenza Artificiale connotati dalla black box, presuppone che sia, in primo luogo, indagato il fenomeno nella sua dimensione reale. La black box, con la quale si intende l'impossibilità di "guardare dentro" il meccanismo di funzionamento del sistema da parte degli stessi programmatori, è peraltro la massima espressione della opacità che caratterizza, con gradi diversi, tutti i modelli algoritmici e connota, in particolare, i sistemi di cd. Deep Learning. Allo stato attuale dello sviluppo tecnologico, pertanto, tali sistemi sembrano essere incompatibili con il principio della trasparenza "rafforzata" che la giurisprudenza amministrativa ha indicato come garanzia minima di legittimità per l'utilizzo di algoritmi nell'attività amministrativa decisionale, anche se privi di "autonomia". Tale principio, inteso come comprensibilità del risultato, è ora espressamente contenuto anche nella proposta di Regolamento della Commissione europea del 21 aprile 2021 come standard di progettazione dei sistemi di I.A.

**Abstract [En]:** The examination of the legal issue of the explicability of administrative decisions, which make use of Artificial Intelligence models characterized by the Black Box, presupposes that, first of all, the phenomenon in its real dimension is investigated. By Black Box we mean the impossibility of "looking inside" the operating mechanism of the system by the programmers themselves, a typical feature, in particular, of the "Deep Learning". At the current state of technological development, therefore, such systems seem to be incompatible with the principle of "enhanced" transparency that administrative jurisprudence has indicated as a minimum guarantee of legitimacy for the use of algorithms in administrative decision-making, even if without "autonomy". This principle, understood as the comprehensibility of the result, is now expressly contained also in the proposal for a Regulation of the European Commission of 21 April 2021 as the design standard for I.A.

**Parole chiave:** Intelligenza artificiale; Black box; algoritmi; amministrazione; tecnologia; deep learning; trasparenza
**Keywords:** Artificial intelligence; Black box; algorithms; administration; technology; deep learning; transparency



## 1. Intelligenza artificiale, black box e le sfide giuridiche delle "tecnologie dirompenti"

Quando una nuova tecnologia emerge dalla società e dall'economia ed approda sulle spiagge del diritto sorgono "*sempre almeno tre categorie di questioni giuridiche: a) quelle già identificate e risolte; b) quelle identificate ma*

* Articolo sottoposto a referaggio. Relazione svoltasi nell'ambito del Corso di formazione per Magistrati amministrativi organizzato dall'Ufficio Studi, massimario e formazione della Giustizia amministrativa "Decisioni automatizzate e pubblica amministrazione. Procedimenti e discrezionalità nel tempo dell'intelligenza artificiale", 8 giugno 2021.

*non risolte; c) quelle non ancora identificate*"[1]. La relazione tra la *black box* e l'esigenza della "esplicabilità" delle decisioni fondate o assistite da sistemi di Intelligenza Artificiale (I.A.) appartiene alla seconda.

La problematica è stata infatti ampiamente individuata, anche prima che il mondo giuridico si occupasse di sistemi di I.A., avendo la giurisprudenza[2] indicato proprio nella esplicabilità, quale declinazione del principio di trasparenza, una delle garanzie minime di legittimità dell'utilizzo di modelli algoritmi per l'adozione di decisioni amministrative; di qualunque algoritmo, anche se "deterministico" e non qualificabile come "intelligente"[3].

La soluzione efficace alla questione è invece ancora lontana, poiché all'orizzonte sembra profilarsi una soluzione variabile e dipendente dal tipo di sistema algoritmico concretamente utilizzato nell'attività amministrativa. Il che conduce ad indagare il fenomeno dell'I.A nella sua dimensione reale, prima che giuridica.

L'esigenza di acquisire innanzi tutto consapevolezza sull'oggetto I.A. nelle sue moltiformi declinazioni ivi compresa la "black box" deriva peraltro da almeno due considerazioni.

La prima è più pragmatica: le nozioni tecniche si trasformeranno a breve in "*definizioni giuridiche*" con le quali è doveroso fare i conti[4]. Invero, all'esito di un lungo, anche se accelerato, percorso delle Istituzioni

---

[1] A. LONGO e G. SCORZA. *Intelligenza artificiale. L'impatto sulle nostre vite, diritti, libertà*, Milano, 2020, p. 206.

[2] Il riferimento è al "filone giurisprudenziale" che si è registrato nel 2019/2020 sulla procedura di mobilità dei docenti, nell'ambito dal piano straordinario di assunzioni di cui alla legge 107/2015 (cd. buona scuola). Le assegnazioni ministeriali dei docenti presso le sedi di servizio erano state gestite mediante un algoritmo – elaborato da fornitori terzi - poi rivelatosi fallace nei risultati, poiché gli interessati erano stati trasferiti in province più lontane da quelle scelte con priorità in sede di partecipazione alla procedura, benché nelle province di elezione i posti fossero rimasti vacanti. L'illegittimità dell'utilizzo dell'algoritmo è stata accertata proprio in considerazione della mancanza di trasparenza che lo connotava, tale da precludere la comprensione del suo meccanismo di funzionamento e conseguentemente della decisione finale fondata su tale elaborazione automatica. Si segnala, in particolare anche per i precedenti ivi citati, Cons. Stato Sez. VI, 13 dicembre 2019, n. 8472, con commento dai R. MATTERA, *Processo - decisioni algoritmiche. Il Consiglio di Stato fissa i limiti*, in *Nuova Giur. Civ.*, 2020, 4, 809; A. MASCOLO, *Gli algoritmi amministrativi: la sfida della comprensibilità*, in *Giornale Dir. Amm.*, 2020, 3, 366; M. TIMO, *Algoritmo - Il procedimento di assunzione del personale scolastico al vaglio del Consiglio di Stato*, in *Giur. It.*, 2020, 5, 1190.

[3] L'algoritmo può essere descritto come una sequenza di istruzioni ordinate in modo preciso e chiaro al fine di trasformare dati di partenza (input) in un qualche risultato (output), il quale va poi "scritto" in un codice con uno dei linguaggi informatici disponibili, perché possa funzionare concretamente in una macchina. Nell'area più generale degli algoritmi, si distinguono, da un lato, "*gli algoritmi deterministici*", in cui tutte le istruzioni, gli input e gli output attesi, nonché i passaggi necessari richiesti per produrre il risultato i criteri sono forniti *ex ante* dal programmatore, i quali presentano una logica lineare e, salvo difetti di progettazione o funzionamento, dovrebbero essere leggibili *ex post*; dall'altro, "*gli algoritmi non deterministici*" in cui la macchina ha un margine di autonomia, con livelli diversi a seconda del modello algoritmico utilizzato. Seconda una delle più accreditate definizioni di I.A. tale "autonomia", che a sua volta presuppone anche una interazione con l'ambiente, è ciò che distingue un sistema "intelligente"; cfr. per una semplificata ricognizione, G. AVANZINI, *Decisioni amministrative e algoritmi informatici. Predeterminazione analisi predittiva e nuove forme di intellegibilità*, Napoli, 2020, pp. 3-13

[4] Ciò, a prescindere dal valore vincolante o meno per l'interprete che si voglia riconoscere alle definizioni legislative: cfr. P. GAGGERO, *A proposito di definizioni legislative*, *Nuova Giur. Civ.*, 2002, 6; R. RORDORF, *Doveri dei soggetti coinvolti nella regolazione della crisi nell'ambito dei principi generali del codice della crisi d'impresa e dell'insolvenza,* in *Fallimento*, 2021, 5, 589, "*è raro trovare nei codici dell'ottocento e della prima metà del novecento, ed in generale nei testi legislativi organici di quel tempo, un corpo di definizioni e di principi generali espressamente destinati a fungere da premessa di tutte le susseguenti disposizioni. Sul valore delle definizioni legislative, per la verità, si discute da sempre, ma è soprattutto negli strumenti normativi sovranazionali - non solo nei trattati e nelle convenzioni, ma*

dell'U.E., la Commissione ha pubblicato in data 21 aprile 2021 la proposta "*Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*" (COM(2021) 206 final) la quale elenca, tra le tecniche ed approcci di I.A., anche i sistemi di "Deep Learning" (cd. apprendimento profondo), tecnicamente connotati dalla "*black box*" e forieri pertanto delle conseguenti problematiche in tema di trasparenza, laddove utilizzati in procedimenti amministrativi.

L'altra considerazione è di più ampio respiro. La conoscenza del "fenomeno" complesso dei sistemi di I.A. costituisce un imprescindibile primo passo per poter affrontare le innumerevoli sfide giuridiche che la loro applicazione nell'attività amministrativa pone; sfide che non è escluso possano condurre ad una rivisitazione di tradizionali categorie giuridiche (atto amministrativo, discrezionalità, ma anche, in particolare sotto il profilo della responsabilità per danni, rapporto di causalità[5]), categorie eventualmente da scomporre e ricomporre per poter rispondere adeguatamente alle nuove esigenze di tutela. Questa prospettiva, peraltro già riscontrata in passate esperienze di "giuridicizzazione" di nuovi interessi economici e sociali[6], appare ancora più concreta in relazione all'I.A. perché l'insieme di tecnologie riassunte nella formula metaforica di I.A.[7] è unanimemente annoverata oggi tra le "*distruptive innovation*"[8],

*anche nei contratti stipulati tra parti aventi nazionalità diverse - che si è andato ben presto diffondendo l'uso di premettere una serie di definizioni volte a chiarire il significato dei termini adoperati poi nel testo ed a vincolare l'interprete a quello specifico significato. Il che agevolmente si spiega tenendo conto proprio del carattere sovranazionale dello strumento normativo e, quindi, della necessità di creare una base di riferimento comune a soggetti che, in quanto provenienti da culture e tradizioni giuridiche diverse, potrebbero altrimenti non bene intendersi sul significato delle parole volte a designare categorie e concetti giuridici*"

[5] Questa prospettiva è stata messa in luce, in particolare con riguardo alla nozione di atto amministrativo e di imputabilità della decisione ad un organo, da I. M. DELGADO, *Automazione, intelligenza artificiale e pubblica amministrazione: vecchie categorie concettuali per nuovi problemi?* In *Istituzioni del federalismo*, 2019, 3.

[6] Basti pensare alla formazione del diritto ambientale e al ripensamento che l'ambiente ha determinato sia in relazione alle categorie dei "*beni pubblici*", sia in relazione alla più basilare categoria di "*situazione giuridica soggettiva*", tanto da indurre autorevole dottrina a ricorrere al diritto romano; cfr. P. MADDALENA, *La scienza del diritto ambientale ed il necessario ricorso alle categorie giuridiche del diritto romano*, in *Rivista quadrimestrale di diritto dell'ambiente*, 2011, 2.

[7] La denominazione è stata coniata, come è noto, dal matematico John McCarthy nel 1956, nell'ambito di un seminario che si svolse presso il Dartmouth College di Hanover, nel New Hampshire. La "congettura" di base su cui si svolsero i due mesi di studio tra autorevoli esponenti della fisica, della matematica, dell'ingegneria informatica, era quella secondo cui "ogni aspetto dell'apprendimento, o qualsiasi altra caratteristica dell'intelligenza umana, possa in linea di principio essere descritto con precisione tale che sia possibile costruire una macchina per simularlo". La natura ingannevole della comparazione tra intelligenza umana ed artificiale è stata poi evidenziata dagli stessi protagonisti del workshop di Dartmouth, comunque oggi identificato come la la data di nascita ufficiale dell'I.A. come autonomo campo disciplinare delle scienze informatiche. A prescindere dalla ancora misteriosa conoscenza dei meccanismi di funzionamento dell'intelligenza umana, una netta demarcazione è stata efficacemente indicata da Roger Penrose, Premio Nobel per la Fisica 2020, secondo cui, a differenza delle macchine, che "*costituiscono un insieme di regole logiche coerenti, ma necessariamente incompleto poiché fondate su assiomi non derivabili automaticamente, la mente umana intuisce la "verità" dei concetti di fondo, è capace di comprendere ciò che conosce, e di farsi una rappresentazione del mondo perché la vera "intelligenza richiede comprensione. E la comprensione richiede consapevolezza*"; capacità non presente ad oggi in nessuna macchina. Cfr. Incontro fra Roger Penrose e Emanuele Severino tenutosi al Centro Congressi Cariplo di Milano il 12 maggio 2018 http://www.vita.it/it/article/2020/10/06/roger-penrose-lintelligenzaartificiale-non-esiste/156896/, visitato il 6 giugno 2021.

[8] L'espressione si deve all'economista Clayton Christensen che la utilizzò per la prima volta in un articolo "*Disruptive technologies: catching the wave*", pubblicato nel 1995 nell'Harvard Business Review. Con essa ci si riferisce a quelle tecnologie che, anche se in una prima fase si presentano in forma primitiva e a costi bassi, sono in breve capaci di far fare al mercato un salto brusco, spazzando via interi sistemi di impresa e forme di organizzazione del lavoro, sostituendoli con nuovi

insieme ad altre "*tecnologie emergenti*"[9] quali la *blockchain*, la realtà aumentata, l'*Internet of Things*, la computazione quantistica, le nanotecnologie; tecnologie che, essendo anche interconnesse tra loro, moltiplicano la loro portata innovativa.

## 2. La black box, ovvero l'impossibilità di "guardare dentro" il meccanismo di funzionamento di alcuni, ma non di tutti, i sistemi di I.A.

Nella nomenclatura attuale, per "*black box*" si intende il massimo livello di opacità che contraddistingue alcuni sistemi di I.A., tale da rendere imperscrutabili, anche agli occhi degli stessi programmatori e sviluppatori, il loro meccanismo di funzionamento e il percorso seguito nella elaborazione degli input (dati) per arrivare ai risultati.

In sostanza, si configura una *black box* quando non è possibile comprendere "l'iter logico" seguito dalla macchina per raggiungere l'obiettivo assegnato; il che, sotto il profilo giuridico, già pare contenere la soluzione, nel senso della incompatibilità, al problema della trasparenza dell'attività amministrativa, laddove un sistema di I.A. che si presenti come una "scatola nera" venga utilizzato nell'ambito del procedimento a supporto della determinazione finale. D'altra parte, anche in senso evocativo, la "black box" richiama immediatamente, e per contrapposizione, la più nota e risalente metafora dell'amministrazione come "casa di vetro"[10] che, applicata alla decisione destinata ad incidere su posizioni giuridiche soggettive, si articola nell'obbligo della motivazione, ovvero proprio nell'obbligo di dar conto dell'iter logico seguito, rendendo palesi le ragioni sottostanti le scelte dell'azione amministrativa.

La *black box,* però, non caratterizza tutti i sistemi di A.I., ma, allo stato dello sviluppo scientifico attuale, in particolare quelli di cd. Deep Learning; cosicché, la consapevolezza del funzionamento dello strumento informatico si rivela necessaria anche per poter operare, a monte, la scelta del modello algoritmico, ovvero del tipo di sistema di I.A. (specie se acquistato presso fornitori terzi)[11], trattandosi di una scelta non neutra ma destinata ad incidere sulla stessa legittimità delle singole decisioni amministrative.

---

modelli di business e nuove professionalità. cfr. C. CHRISTENSEN, *Il dilemma dell'innovatore: come le nuove tecnologie possono estromettere dal mercato le grandi aziende*, 1997.

[9] Alle "*tecnologie emergenti*" si riferisce testualmente, nell'ordinamento giuridico italiano, l'art. 8 comma 2 lett. b) del D.L. 1 marzo 2021 n. 22, convertito dalla legge 22 aprile 2021, n. 55; locuzione esemplificata con il richiamo espresso all'I.A., alla *blockchain* e all'Internet delle cose (IoT), indicate dal legislatore tra le priorità di intervento del neoistituito Comitato interministeriale per la transizione digitale (CITD).

[10] L'espressione, che originariamente fu l'auspicio invocato da Filippo Turati alla Camera dei Deputati nel 1908 "*Dove un superiore, pubblico interesse non imponga un momentaneo segreto, la casa dell'amministrazione dovrebbe essere di vetro*", è stata ultimo citata dall' Ad. Plen. 10/2020 in materia di accesso agli atti di gara, proprio specificando che "*Il principio di trasparenza (…) costituisce anche un caposaldo del principio di buon funzionamento della pubblica amministrazione, quale "casa di vetro" improntata ad imparzialità, intesa non quale mera conoscibilità, garantita dalla pubblicità, ma anche come intelligibilità dei processi decisionali e assenza di corruzione*" Cons. Stato, Ad. Plen. 2 aprile 2020, n. 10; cfr. L. GIAMPIETRO, *La P.A. nella "casa di vetro": i concorrenti "diritti di accesso" in materia di appalti pubblici,* in *Ambiente e sviluppo, 2020, 8-9, 671*

[11] E ciò, sia se la scelta deve essere effettuata nell'ambito della valutazione comparativa delle diverse soluzioni disponibili imposta dall'art. 68 del d.lgs. 7 marzo 2005, 82 (CAD); sia, e a maggior ragione, nell'ipotesi in cui l'esigenza non possa

Il richiamo ai sistemi di Deep Learning - e quindi della black box che li connota - è espressamente contenuto nella articolata definizione legislativa del sistema di I.A. che emerge dall'art. 3 punto 1 della proposta di Regolamento della Commissione dell'U.E. del 21 aprile 2021[12] secondo la quale esso consiste in un "*software sviluppato con una o più delle tecniche e degli approcci elencati nell'allegato I e che può, per una data serie di obiettivi definiti dall'uomo, generare risultati quali contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono*". La definizione, da un lato, mette in evidenza l'interazione con l'ambiente e la funzionalizzazione dei sistemi di I.A. a perseguire specifici compiti ("*obiettivi definiti*") quali caratteristiche generali di ogni sistema di I.A., ma, dall'altro, rinvia all'allegato 1, rubricato "*Tecniche e approcci di intelligenza artificiale*", per descrivere le diverse tipologie disponibili, ovvero i diversi modelli algoritmici dei sistemi di I.A.[13], richiamando -in particolare nelle lett. a) e b)- la distinzione tra le due grandi "famiglie" in cui essi possono suddividersi[14]: sistemi non simbolici, ovvero di apprendimento automatico (Machine Learning, di cui il Deep Learning è un sottoinsieme) e sistemi simbolici.

essere soddisfatta ricorrendo a soluzioni già disponibili sul mercato l'amministrazione ricorra ad esempio al partenariato per l'innovazione ai sensi dell'art. 65 del D.Lgs. 18 aprile 2016, n. 50.

[12] L'esigenza di definire l'I.A., per individuare il perimetro delle regole ad essa applicabili, permea tutto il "fermento" normativo delle Istituzioni dell'U.E. Da ultimo, nel Libro Bianco del 19 febbraio 2020, la Commissione definisce l'IA come "*un insieme di tecnologie che combina dati, algoritmi e potenza di calcolo*", facendo emergere la stretta connessione, che ha determinato gli sviluppi più recenti, tra I.A. come è intesa oggi e l'ampia disponibilità di dati (cd. Big Data): "*i progressi compiuti nell'ambito del calcolo e la crescente disponibilità di dati sono pertanto fattori determinanti per l'attuale crescita dell'IA" (COM(2020)65 final);* Anche il Parlamento Europeo si è cimentato con la definizione giuridica di I.A.; invero, la Risoluzione del Parlamento europeo del 16 febbraio 2017 prende l'abbrivio citando miti epici e romanzi di fantascienza *"dal mostro di Frankestein ideato da Mary Shelley al mito classico di Pigmalione, passando per la storia del Golem di Praga e il robot di Karel Capeck, che ha coniato la parola, gli essere umani hanno fantasticato sulla possibilità di costruire macchine intelligenti, spesso androidi con caratteristiche umane*" e con essa il Parlamento sollecita la Commissione ad adottare una nozione che tenga conto dell'autonomia e dell'adattività dei sistemi di intelligenza artificiale *(Risoluzione del Parlamento europeo del 16 febbraio 2017 recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_IT.html).* Nella più recente Risoluzione del Parlamento europeo del 20 gennaio 2021 *"Intelligenza artificiale: questioni relative all'interpretazione e applicazione del diritto internazionale",* con la quale viene sollecitata anche una regolazione anche dei preoccupanti usi militari (settore oggetto di particolare attenzione negli Stati Uniti, ma rimasto fuori sia dal Libro Bianco del 19 febbraio 2020, che dalla proposta di regolamento della Commissione del 21 aprile 2021), si propone l'adozione di una definizione più di I.A., fondata sulla duplice caratteristica della simulazione del comportamento umano e di "un certo grado di autonomia"*: "sistema basato su software o integrato in dispositivi hardware che mostra un comportamento che simula l'intelligenza, tra l'altro raccogliendo e trattando dati, analizzando e interpretando il proprio ambiente e intraprendendo azioni, con un certo grado di autonomia, per raggiungere obiettivi specifici* (*P9_TA-PROV(2021)0009).*

[13] Allegato 1: *"a) approcci all'apprendimento automatico, compreso l'apprendimento supervisionato, non supervisionato e di rinforzo, utilizzando un'ampia varietà di metodi compreso l'apprendimento profondo; (b) approcci basati sulla logica e sulla conoscenza, compresa la rappresentazione della conoscenza, la programmazione (logica) induttiva, le basi della conoscenza, i motori inferenziali e deduttivi, il ragionamento (simbolico) e i sistemi esperti; (c) Approcci statistici, stima bayesiana, metodi di ricerca e ottimizzazione".* La scelta di articolare la definizione di I.A. su due livelli normativi è determinata dalla ovvia constatazione che lo sviluppo tecnologico in questo campo è così veloce che risulta inefficace cristallizzare le "tecniche e gli approcci" in una disposizione normativa, la cui modifica presuppone tempi incompatibili con la necessità di aggiornamento. L'art. 4 della proposta di Regolamento attribuisce infatti alla Commissione il potere di adottare atti delegati con i quali procedere alla revisione dell'elencazione contenuta nell'allegato 1, tentando così di semplificare la difficile opera di inseguimento da parte del legislatore del nuovo che avanza.

[14] Si tralascia in questa sede, anche per esigenze di sinteticità, la descrizione dei modelli "statistici" indicati nell'allegato 1 alla lett. c)

Questi ultimi hanno caratterizzato il settore dai suoi albori fino agli anni '80 (tanto da essere oggi etichettati come *GOFAI, Good Old-Fashioned Artificiale Intelligence*): essi sono caratterizzati dalla rappresentazione simbolica della conoscenza e da istruzioni predefinite prestabilite dai programmatori. Tra questi, quelli più diffusi sono i cd. "*sistemi esperti*" in cui, in un ambito predefinito (es. in uno specifico settore medico), viene fornita alla macchina una conoscenza di base corrispondente appunto a quella di "*un esperto del settore*". Ciò che consente di classificare tali modelli algoritmici come "intelligenti" (così distinguendoli dagli algoritmi "deterministici" privi di autonomia) è che essi mostrano un'abilità umana, simile alla capacità logico-deduttiva di ragionamento; sono cioè capaci di dedurre, dai dati forniti, informazioni nuove, eventualmente producendo anche risultati in termini di probabilità se i dati sono incompleti. I sistemi simbolici hanno pertanto un grande vantaggio, in termini di "compatibilità" con le regole e con il ragionamento giuridico, articolato anch'esso su logiche argomentative presupposti/conseguenze o logiche inferenziali cause/effetto: il loro meccanismo è trasparente, poiché è possibile ripercorrere all'indietro il processo e comprenderne il meccanismo di funzionamento.

Dagli anni '90 in poi, ovvero in concomitanza con la sempre più diffusa digitalizzazione della realtà, determinata dalla esplosione di Internet e dei social web, i modelli di I.A. su cui si è concentrata la ricerca, e che pervadono anche le applicazioni oramai entrate nella vita quotidiana, vengono annoverati nella diversa categoria del Machine Learning (cd. apprendimento automatico). La filosofia di fondo di questa nuova generazione di I.A. è completamente diversa, poiché poggia sulla considerazione che la "conoscenza" è quella che si acquisisce con l'esperienza e l'apprendimento: al sistema di Machine Learning vengono pertanto fornite, non istruzioni, ma quantità enormi di dati, quali "esempi" sui quali la macchina, supportata da una elevata capacità di calcolo, viene addestrata al fine di perseguire l'obiettivo assegnato, prima di entrare a regime. Grazie alla quantità di dati (Big Data) elaborati durante la fase di addestramento, gli obiettivi assegnati (es. riconoscimento immagini, riconoscimento vocale, interpretazione linguaggio naturale) vengono conseguiti con affidabilità sempre più elevate, poiché il sistema individua, sulla base dei dati immessi correlazioni nascoste fornendo risposte non spiegabili secondo logiche argomentative.

Nell'ambito dell'insieme di sistemi di Machine Learning si delinea, come emerge anche dall'ultima parte della lett. a) dell'allegato 1 della proposta di Regolamento ("*compreso il Deep Learning*"), lo specifico campo applicativo del *Deep Learning*[15] che, per la complessità dei modelli matematici utilizzati (cd. "*reti neurali*

[15] I sistemi di Deep Learning utilizzano un'architettura di modelli matematici ispirata alle reti neurali biologiche: le cd. reti neurali artificiali. Tale modello è costituito da un gruppo di interconnessioni di informazioni (si parla infatti di approccio di "connessionismo" al calcolo, contrapposto all'approccio simbolista): gli input trasmettono i segnali, ad una potenza ovviamente incomparabile con quella dei neuroni biologici, ai diversi nodi che costituiscono una rete complessa (deep) e nel corso dell'apprendimento, i "pesi" di ciascun nodo vengono continuamente riparametrati, in un percorso non lineare e multistrato la cui ricostruzione però sfugge alla comprensione umana.

*artificiali*"), è caratterizzato appunto dalla *black box*, ovvero dalla pratica impossibilità di comprendere "come" e "perché", sulla base del set di dati elaborati, il sistema sia giunto a certi risultati; impossibilità dovuta alla complessità dei calcoli e alla natura "non lineare" del processo matematico di elaborazione e sulla quale vi è una convergenza di opinioni tra tutti gli esperti del settore[16].

I sistemi di Deep Learning portano, pertanto, alle estreme conseguenze il tendenziale disallineamento tra l'elaborazione informatica complessa di dati e il ragionamento giuridico, poiché in essi il "modello" viene costruito a posteriori dalla macchina che, quasi invertendo il metodo scientifico moderno che dalla ipotesi giunge alle tesi, individua, nella mole di dati a disposizioni, schemi ricorrenti, *pattern*, regolarità statistiche; la decisione è pertanto fondata sui dati e non vi è una spiegazione "ragionevole" del risultato perché il sistema non è geneticamente costruito per dare motivazioni.

La decisione fondata su regole giuridiche segue invece un modello normativo prestabilito che è rappresentato a monte dalla fattispecie astratta – la quale peraltro presenta quasi sempre profili di ambiguità e viene interpretata con risultati non univoci – e che viene applicato nel "caso concreto" sulla base delle informazioni raccolte, al fine di adottare una decisione che sia ragionevole e intellegibile[17].

---

[16] Tanto da aver fatto emergere uno specifico campo di ricerca volto ad "aprire" la black box dei sistemi di Deep Learning: la cd. *eXplainable A.I. cd. XAI*. Per una ricognizione sia delle architetture dei modelli complessi sia delle applicazioni più diffuse di Deep Learning, cfr. R. CUCCHIARA, *L'intelligenza non è artificiale*, Milano, 2021, pag. 67 e ss.

[17] Si tratta pertanto di "logiche" diverse, il cui dialogo non è scontato, anche perché il "*dato*" (ovvero gli esempi che sono forniti al sistema di I.A.) non è la realtà, ma solo la sua "rappresentazione", mentre la realtà è per sua natura irriducibile ad una rappresentazione, come un territorio non è riducibile alla sua mappa; nel mondo del diritto, in cui il modello normativo precede logicamente il fatto concreto con l'ambizione di "*dominare l'im-prevedibile* (…) *il caso è un evento im-prevedibile, ma riconoscibile. L'imprevedibile ha dentro di sé l'idea del divenire continuo, di uno scorrere di atti umani e fatti naturali, che non siamo in grado di determinare con necessità né di prevedere con qualche certezza. L'imprevedibile irrompe nella vita, spesso non atteso, non augurato, non voluto (…). La parola "caso" indica propriamente la duplice natura del fatto concreto: da un lato, l'oggettiva imprevedibilità del suo accadere (..); dall'altro, l'oggettiva conformità dell'accaduto al modello normativo, di cui appunto viene a costituire uno degli innumerevoli esempi*", N. IRTI, *Un diritto incalcolabile*, Torino, p. 19. Sul punto, cfr. Kevin Ashley, *Artificial Intelligence and Legal Analiytics: new tools for law pratics in the Digital Age*, Cambridge, 2017, "*le previsioni di Machine Learning sono basate sui dati. A volte i dati contengono caratteristiche che, per motivi spuri, come la coincidenza o la selezione parziale sono associati ai risultati dei casi in una particolare raccolta. Sebbene le regole indotte dalla macchina possano portare a previsioni accurate, esse non si riferiscono all'esperienza umana e potrebbero non essere così comprensibili per l'uomo come le regole costruire manualmente da un esperto. Poiché le regole che l'algoritmo di Machine Learning inferisce non riflettono necessariamente le conoscenze o le competenze giuridiche potrebbero non corrispondere a criteri di ragionevolezza di un esperto umano*" (riportato da A. SANTOSUOSSO, cit., p. 99). Va osservato che la prospettiva della divergenza generale tra decisioni fondate su dati e decisioni fondate su regole giuridiche è un punto realmente critico nelle applicazioni di tecnologie di A.I. applicate al diritto, il cui approfondimento presupporrebbe che sia innanzi tutto chiarito *a priori* il termine di paragone, ovvero che sia chiaro quale sia l'effettivo contenuto della "motivazione" rispetto al processo decisionale concreto, trattandosi di elementi che non necessariamente coincidono neanche nei meccanismi cognitivi umani, restando margini "oscuri", sfumature, passaggi logici mentali intuitivi, i quali restano oscuri alla pari di una *black box,* senza che ciò infici la legittimità in termini di trasparenza della determinazione. Sotto questo profilo, è stato osservato come, per "*ridimensionare l'ansia da spiegabilità della decisione che sembra pervadere alcune posizioni contrarie pregiudizialmente all'uso di I.A. nel decidere (..) si può cominciare ad accettare che* (anche) *gli umani hanno i loro pregiudizi e un loro modo unico e non conoscibile di prendere decisioni*". Per un richiamo al tema dei *bias c*ognitivi nel momento decisionale che potrebbero essere conosciuti e "sfruttati" anche dal legislatore al fine di delineare sistemi di regole efficienti ed efficaci, secondo l'approccio della "*behavioural regulation*", cfr. parere del Consiglio di Stato, Sezione Consultiva per gli Atti Normativi, n. 1458 del 19 giugno 2017; cfr. in generale, D. KAHNEMAN, *Pensieri lenti e veloci*, Milano, 2017.

### 3. La *black box* nell'ambito della più generale opacità dei sistemi algoritmici e il baluardo della "trasparenza rafforzata"

La *black box* dei sistemi di Deep Learning è pertanto una opacità assoluta.

Ma l'opacità caratterizza, in forme e gradi diversi, tutti i modelli algoritmici, anche quelli deterministici (non di I.A.) e pertanto fa emergere per tutti la questione della compatibilità del sistema con la doverosa trasparenza dell'azione amministrativa e con la esplicabilità delle determinazioni finali, specie se adottate nell'esercizio di poteri autoritativi.

Diversi sono i fattori che rendono oscuri i modelli algoritmici: a) in primo luogo essi sono appannaggio delle scienze ingegneristiche, informatiche e logico-matematiche e sono trascritti, per poter essere operativi ed elaborati da una macchina, in "codici sorgente" utilizzando, non il linguaggio naturale con cui si esprimono le regole e l'argomentazione giuridiche, ma in uno dei 2.500 linguaggi informatici disponibili; b) gli algoritmi più performanti tecnologicamente sono inoltre anche entità commerciali, soggette a leggi industriali e proprietà intellettuali, protetti dalle aziende produttrici sia agli occhi dei *competitor* che dagli eventuali malintenzionati volti a "sabotarli"[18]; c) in ogni caso, un software, ancora di più quelli tecnologicamente evoluti di I.A., è sempre in continua evoluzione, è per definizione "instabile", soggetto a continui aggiornamenti, versioni, integrazioni, pena il malfunzionamento o il non funzionamento totale[19].

A fronte di tale opacità a "geometria variabile" e a prescindere dalla natura intelligente e dal livello di autonomia degli algoritmi, è stato merito della giurisprudenza del Consiglio di Stato, a monte, risolvere il problema a monte della conformità del loro utilizzo rispetto al principio di legalità,[20] ancorandosi ai

[18] Invero la brevettabilità dei soli modelli algoritmici, che poi consistono in modelli matematici, prima che essi siano trascritti in "codici sorgente" è oggetto di ampia discussione, in tutti i Paesi dell'Unione Europea. cfr. T. FAELLI, Le innovazioni in materia di tecnologia blockchain tra diritto dei brevetti e diritto d'autore, in Dir. Industriale, 2020, 2, p.172; S. MAGELLI Le nuove tecnologie nella giurisprudenza, in Dir. Industriale, 2020, 2, p. 199; nell'ordinamento giuridico nazionale, l'art. 45 comma 2 del D.Lgs. 10 febbraio 2005, n. 30 cd. Codice della proprietà industriale, espressamente prevede che "non sono considerate invenzioni", suscettibili di essere oggetto di brevetto, tra l'altro "i metodi matematici" (lett.a) e i "programmi per elaboratore" (lett. b).

[19] "*La comprensione di un fenomeno che cambia velocemente può rendere tutto molto più oscuro e impenetrabile: quello che abbiamo capito oggi del software potrebbe non essere così illuminante domani e, inoltre, spesso non riusciamo più neanche a leggere i programmi software usati nel passato*", C. ACCOTO, *Il mondo dato. Cinque brevi lezioni di filosofia digitale*, Milano, 2017, p. 11. La traduzione filosofica di questa nuova ontologia digitale in cui siamo tutte creature dell'Aggiornamento, in cui "Essere" è "Essere aggiornati" assume rilievo anche sul lato della formazione e dell'apprendimento umano delle nuove tecnologie; emerge cioè l'importanza anche della capacità di disimparare e re-imparare rapidamente da parte dei consumatori e dei professionisti: "*a fronte di questo stato di cambiamento continuo, sapere disimparare (unlearning) sarà una competenza professionale e personale importante almeno quanto saper apprendere rapidamente e saper ricercare in rete conoscenza e informazione*" (C. ACCOTO, cit., pag. 20)

[20] Soluzione ora peraltro pienamente convalidata dalla novella dell'art. 3-bis della legge 241/90 per effetto dell'art. 12 del D.L. 16 luglio 2020, n. 76 convertito, con modificazioni, dalla L. 11 settembre 2020, n. 120. In particolare, all'esito di tale novella, la citata norma prescrive oggi che "*Per conseguire maggiore efficienza nella loro attività, le amministrazioni pubbliche agiscono mediante strumenti informatici e telematici, nei rapporti interni, tra le diverse amministrazioni e tra queste e i privati*" segnando così un doppio cambio di passo: da un lato, la trasposizione nella sede naturale della legge sul procedimento amministrativo di un canone già evincibile dall'interpretazione sistematica dell'art. 12 del l'art. 12, comma 1, del CAD ("*Le Pubbliche Amministrazioni nell'organizzare autonomamente la propria attività utilizzano le tecnologie dell'informazione e della*

principi costituzionali e sovranazionali di "buona amministrazione" e, a valle, indicare "*gli elementi di minima garanzia per ogni ipotesi di utilizzo di algoritmi in sede decisoria pubblica",* costituiti, in particolare, dalla *"a) piena conoscibilità a monte del modulo utilizzato e dei criteri applicati"* e dalla "*b) imputabilità della decisione all'organo titolare del potere, il quale deve poter svolgere la necessaria verifica di logicità e legittimità della scelta e degli esiti affidati all'algoritmo*", secondo una "*declinazione rafforzata del principio di trasparenza*[21]*, che implica anche quello della piena conoscibilità di una regola espressa in un linguaggio differente da quello giuridico.*"[22].

Posto come grimaldello per superare l'opacità degli algoritmi e assicurare la legittimità almeno sotto questo profilo della decisione amministrativa che se ne avvalga, il principio di trasparenza si articola, peraltro, in due livelli di garanzia: da un lato, nella necessaria consapevolezza che la decisione, nel caso concreto, sia fondata o assistita da sistemi automatizzati (cd. *principio di conoscibilità*); dall'altro, che siano fornite al destinatario informazioni utili a "*comprendere"* le modalità di funzionamento del sistema, in modo da poterne verificare i criteri, i presupposti, gli esiti, in sostanza la "*logica*" utilizzata e, quindi, la sua conformità alla fattispecie astratta (al modello pre-dato) che attribuisce il potere. Insieme, la conoscibilità astratta della natura "robotica" della decisione e la comprensibilità del suo contenuto confluiscono nel principio composito della "*esplicabilità*"[23].

Ciò che più rileva nel paradigma digitale della trasparenza è però la tassonomia tra i due requisiti di legittimità sopra indicati: quello della esplicabilità e della necessaria imputabilità della decisione ad una persona fisica. In mancanza di spiegabilità, anche il principio di responsabilità ultima della decisione resta svuotato, poiché il maccanismo di funzionamento del sistema algoritmico deve essere comprensibile, in primo luogo, per il titolare dell'organo che assume la decisione (oltre che per tutti coloro che sono a vario titolo coinvolti nel procedimento decisionale, specie se la decisione è pluristrutturata, come avviene ad

---

*comunicazione per la realizzazione degli obiettivi di efficienza, efficacia, economicità, imparzialità, trasparenza, semplificazione e partecipazione*") e dell'art. 41 co. 1 "*Le Pubbliche Amministrazioni gestiscono i procedimenti amministrativi utilizzando le tecnologie dell'informazione e della comunicazione*"; dall'altro, la sostituzione del modello precedente, incentivante e meramente programmatico (l'art. 3-bis previgente prevedeva che "*Per conseguire maggiore efficienza nella loro attività, le Amministrazioni Pubbliche incentivano l'uso della telematica, nei rapporti interni, tra le diverse amministrazioni e tra queste e i privati*") con un modello precettivo, tale da qualificare come doveroso il ricorso ad un metodo algoritmico qualora, tenuto conto dell'obiettivo assegnato all'azione amministrativa e della disciplina specifica, esso dia maggiori garanzie di efficienza rispetto all'opzione-zero.

[21] I riferimenti normativi del principio di trasparenza rafforzata per i sistemi algoritmici sono stati individuati, sul piano interno, nell'art. 97 Cost, quale garanzia di buon funzionamento dell'amministrazione; ma anche in riferimenti normative sovranazionale, ovvero, quanto al diritto primario dell'Unione Europea, nell'art. 41 della Carta Europea dei Diritti Fondamentali ("*Right to a good administration*") e, comunque, nelle plurime disposizioni del Regolamento UE n. 679/2016 del Parlamento Europeo e del Consiglio del 27 aprile 2016 sulla protezione dei dati personali (GDPR) che riconoscono il diritto a conoscere dell'esistenza di processi decisionali automatizzati che abbiano effetti giuridici sui destinatari e - nelle ipotesi in cui operino le deroghe al generale divieto di una decisione che sia unicamente fondata sul trattamento automatizzato di dati di cui all'art. 22 par. 2 - quello a ricevere "*informazioni significative sulla logica utilizzata*"; Cfr. in particolare, artt. 13, par. 2, lett. f, 14, par. 2, lett. g e 15, par. 1, lett. h.

[22] Cons. Stato Sez. VI, 13 dicembre 2019, n. 8472; cfr. *supra* nota 2.

[23] A. SIMONCINI, *Diritto costituzionale e decisioni algoritmiche,* in *Il ragionamento giuridico nell'era dell'intelligenza artificiale*, S. DORIGO (a cura di), Pisa, 2020, p. 53

esempio nel caso della conferenza di servizi ex art. 14 e ss. della Legge 241/90) e valida la "pre-decisione", il risultato dell'istruttoria algoritmica, la raccomandazione, la proposta offerta dal modello matematico. Questo legame logico tra i due elementi di garanzia richiama invero un'ulteriore problematica, che è quella della cd. "cattura del decisore", ovvero della invisibile, ma inevitabile depauperazione della capacità decisionale umana, via via che l'elaborazione automatica diventa più affidabile ed efficiente ad appare pertanto anche "oggettiva ed imparziale"; ovvero il rischio concreto del cd. *automation bias*, da tenere in considerazione nella concreta operatività dei meccanismi decisionali[24].

Ma la priorità logica del principio di esplicabilità sussiste anche rispetto al divieto di discriminazione algoritmica, che è il pericolo avvertito come più allarmante nell'ambito dei sistemi di I.A., anche per effetto dei noti casi giudiziari che oramai fanno parte della letteratura scientifica[25]. I sistemi di I.A., elaborando e proiettando nel futuro dati immessi in quantità notevoli che provengono dal passato, tendono a automatizzare e cristallizzare le discriminazioni che i dati nascondono, in termini di genere, razza, reddito, provenienza geografica. Senza la spiegazione del meccanismo di funzionamento non è possibile comprendere dove si annida il risultato discriminatorio, se nel sistema algoritmico o se nel set di dati utilizzati, e non è possibile pertanto né prevenire il rischio discriminatorio né correre ai ripari in caso violazione del divieto di discriminazioni[26].

---

[24] "*Il principio di non esclusività sembra non riconoscere l'evidente forza "pratica" di qualsiasi automatismo valutativo che, da un lato, solleva il decisore dal burden of motivation, dal peso dell'esame e della motivazione; dall'altro gli consente di "qualificare" la propria decisione con un crisma di "scientificità" ovvero "neutralità" che oggi circonda la valutazione algoritmica e le conferisce una peculiare – quanto infondata – "autorità"*", Andrea Simoncini, cit., p. 56.; cfr. P. GALLINA, *La mente liquida. Come le macchine condizionano, modificano o potenziano il cervello*, Bari, 2019

[25] Un approfondito resoconto del cd. caso Eric Loomis – citato anche dalla sentenza del Consiglio di Stato 8472/2019- e del risultato discriminatorio, per ragioni razziali, dell'algoritmo Compas utilizzato dalle Corti penali americane per quantificare il rischio di recidiva è riportato in A. LONGO e G. SCORZA, cit., p. 217 e ss. Più recentemente, e con riguardo all'Unione Europea, un analogo caso di sistema algoritmico risultato affetto da gravi *bias* discriminatori è stato quello dell'applicazione SyRi (*System Risk Indication*) usata dall'amministrazione olandese dal 2014 fino alla sua sospensione da parte del Tribunale dell'Aja nel febbraio 2020. Si trattava di un sistema che, al fine di valutare l'attitudine a commettere frodi o abusi da parte dei beneficiari di sussidi statali, attingeva dati sensibili da 17 banche-dati, attribuendo all'interessato un "punteggio di rischio". I dati elaborati provenivano però solo da quartieri periferici delle città di Rotterdam, Eindhoven e Haarlem, abitati da residenti con basso reddito, migranti e appartenenti a minoranze etniche. In ragione della "cristallizzazione" della discriminazione che tale set di dati aveva determinato a causa della loro conformazione *ex ante,* una coalizione di associazioni per la tutela dei diritti digitali ha adito il tribunale distrettuale dell'Aia per verificare la legittimità dell'utilizzo di SyRi in relazione alla tutela del diritto alla privacy di ogni cittadino. All'esito del processo, il programma è stato ritenuto troppo invasivo sulla vita delle persone, in contrasto con l'art. 8 CEDU, nonché non conforme ai principi di trasparenza e di minimizzazione di cui al Regolamento UE 2016/679. Tale vicenda ha peraltro determinato anche una crisi politica, poi sfociata nelle dimissioni del Primo Ministro olandese a gennaio del 2021.

[26] La connessione logica tra comprensibilità e gli altri requisiti di legittimità della decisione "automatizzata" è stata da ultimo affermata anche dalla Cassazione civile che, con l'ordinanza n. 14381 del 25 maggio 2021, ha affermato il principio secondo cui, in tema di trattamento dei dati personali, il consenso valido presuppone la consapevolezza di chi lo esprime che a sua volta deve comprendere la conoscenza dello "schema esecutivo dell'algoritmo" (nel caso di specie, posto a base di una piattaforma web che elaborava requisiti reputazionali di singole persone fisiche o giuridiche secondo punteggi di affidabilità): "*non può logicamente affermarsi che l'adesione alla piattaforma da parte dei consociati comprenda anche l'accettazione di un sistema automatizzato, che si avvale di un algoritmo, per la valutazione oggettiva di dati personali, laddove non siano resi conoscibili lo schema esecutivo in cui l'algoritmo si esprime e gli elementi all'uopo considerati*"

## 4. Trasparenza ed esplicabilità *by design* secondo la proposta del Regolamento del 21 aprile 2021: un nuovo approccio culturale

Le istituzioni europee hanno manifestato un crescente fervore per l'intelligenza artificiale, spinte anche dall'analogo fermento mostrato da Stati Uniti[27] e Cina[28], le cui iniziative governative mirano a contendersi la *leadership* mondiale nel settore.

L'approccio dell'Unione Europea nei confronti delle tecnologie emergenti si muove in particolare su un binario[29], essendo teso, da un lato, a recuperare terreno rispetto al *gap* che ancora separa l'Europa da Stati Uniti e Cina, specie in termini di investimenti pubblici nella ricerca; dall'altro, a proiettare tale sviluppo nell'alveo dei valori e diritti fondamentali dell'ordinamento giuridico europeo, perseguendo l'obiettivo di incentivare una "*intelligenza artificiale antropocentrica*"[30], creata dall'uomo, per l'uomo e destinata a restare, per essere "sostenibile", sotto la sua sorveglianza.

---

[27] Con la legge di Bilancio della Difesa per il 2019 (*NDAA for fiscal Year 2019*), è stata istituita negli Stati Uniti la *National Security Commission on Artificial Intelligence*, composta da quindici membri designati dalle commissioni parlamentari e dai Ministri di Difesa e Commercio. La Commissione ha il compito di elaborare proposte per una strategia efficace nel campo dell'IA, con prevalente attenzione al campo militare e al tema della sicurezza. Il rapporto finale del 2021 delinea un quadro preoccupante per l'accelerazione impressa agli investimenti e alle politiche di supporto alle imprese e alla ricerca di talenti da parte della Cina, reclamando una maggiore leadership della Casa Bianca e una coalizione con le imprese, gli enti di ricerca, i partner internazionali e sancendo la mancata preparazione degli Stati Uniti all'impatto anche geopolitico delle nuove tecnologie. Le conclusioni prendono l'abbrivo da una considerazione sull'efficacia dirompente della nuova tecnologia "*L'intelligenza artificiale non è una singola tecnologia svolta, come un bombardiere stealth con ala di pipistrello. La corsa all'AI la supremazia non è come la corsa allo spazio verso la luna. L'intelligenza artificiale non è pari paragonabile a una tecnologia di uso generale come l'elettricità. Tuttavia, ciò che Thomas Edison ha detto dell'elettricità racchiude il futuro dell'intelligenza artificiale: "E' il campo dei campi… custodisce i segreti che riorganizzeranno la vita del mondo." La sbalorditiva valutazione di Edison veniva dall'umiltà. Ciò che lui aveva scoperto era «molto poco in confronto alle possibilità che appaiono*." Tra le raccomandazioni finali, uno spazio è riservato anche al tema della trasparenza, nel quale si prende atto del fenomeno della "black box" che connota alcune applicazioni di Machine Learning: "*Migliorare la trasparenza pubblica su come il governo utilizza l'IA. C'è una mancanza di trasparenza nelle politiche e procedure delle agenzie e nell'accuratezza dei sistemi di intelligenza artificiale che possono avere un impatto sulle libertà civili. La natura di "black box" di alcuni sistemi ML non fa che aumentare questa opacità. Una maggiore trasparenza potrebbe contribuire ad alleviare le preoccupazioni del pubblico. Certo, in certi* contesti operativi, *in particolare per l'intelligence e le forze dell'ordine, la segretezza è essenziale per la* missione. *Tuttavia, i meccanismi di trasparenza esistenti potrebbero essere utilizzati più efficacemente e, in alcuni casi, rivisti*" https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1., visitato il 10 giugno 2021.

[28] Per una ricognizione degli interventi governativi e normativi in Cina, in relazione agli sviluppi della tecnologia informatica e ai suoi impatti sulla tutela dei "diritti di libertà", A. MALASCHINI, *Diritti di libertà, controllo sociale e tecnologia informatica: gli sviluppi recenti del modello cinese*, in *Le sfide del diritto transnazionale. Percorsi costituzionali*, 2-3, 2018.

[29] A. ADINOLFI, *L'Unione europea dinnanzi allo sviluppo dell'intelligenza artificiale: la costruzione di uno schema di regolamentazione europeo tra mercato unico digitale e tutela dei diritti fondamentali*, in (a cura di S. DORIGO), Il *ragionamento giuridico nell'era dell'intelligenza artificiale*, Pisa, 2020, p.17.

[30] L'obiettivo di una Intelligenza artificiale giuridico-sostenibile emerge fin dai primi documenti elaborati in materia dalle Istituzioni europee: è indicato nella comunicazione del 25 aprile 2018 "*Strategia per l'I.A*" COM(2018) 237 final con cui si avvia il percorso che, attraverso l'elaborazione delle Linee Guida Etiche -predisposte *dall'High Level Expert Group on Artificiale Intelligence* e confluite nella Comunicazione "*Creare fiducia nell'intelligenza artificiale antropocentrica*" dell'8 aprile 2019 (COM(2019) 168 final) - ha portato, dapprima al "*Libro bianco sull'intelligenza artificiale Un approccio europeo all'eccellenza e alla fiducia* (19.2.2020 COM(2020)65 e, da ultimo, alla proposta presentata al Parlamento e al Consiglio europeo il 21 marzo 2021 "*Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*" (COM(2021) 206 final), già citata.

La spiegabilità dei sistemi di intelligenza artificiale è un elemento chiave di questa architettura volta a realizzare una I.A. "sostenibile".

Essa era già stata indicata come uno dei sette requisiti fondamentali per un'IA affidabile contenuti nelle Linee Guida Etiche del 2019 (intervento e sorveglianza umani, robustezza tecnica e sicurezza, riservatezza e *governance* dei dati, trasparenza, diversità non discriminazione ed equità, benessere sociale e ambientale, *accountability)*, i quali costituiscono i pilastri sui cui è stata elaborata la proposta di Regolamento del 21 aprile 2021. In particolare, in relazione al requisito della trasparenza, la Commissione, facendo proprie le conclusioni del Gruppo di Esperti nominati ad aprile del 2018 confluite nelle Linee Guida, aveva indicato tre livelli di garanzia minimi di trasparenza: a) la tracciabilità, intesa come la documentazione e registrazione delle decisioni adottate dai sistemi di intelligenza artificiale e dell'intero processo di elaborazione della decisone; b) la spiegabilità "*per quanto possibile*" del processo decisionale degli algoritmi, adattata alle persone coinvolte; c) la comunicazione, adeguata al caso concreto, delle capacità e dei limiti del sistema di I.A., al fine di garantire la consapevolezza da parte degli utenti che essi stanno interagendo con un sistema di I.A.[31]

Nella proposta di Regolamento citata, la Commissione propone ora una più articolata applicazione del principio di trasparenza intesa come "esplicabilità", in linea con il principio di proporzionalità sotteso alla classificazione delle "pratiche" di I.A. a seconda del diverso livello di "rischio"[32].

---

[31] "*La tracciabilità dei sistemi di IA dovrebbe essere garantita: è importante registrare e documentare sia le decisioni adottate dai sistemi sia l'intero processo che ha prodotto le decisioni, comprese una descrizione della raccolta e dell'etichettatura dei dati e una descrizione dell'algoritmo utilizzato. In questo contesto dovrebbe essere prevista, per quanto possibile, la spiegabilità del processo decisionale degli algoritmi, adattata alle persone coinvolte. È opportuno proseguire la ricerca in corso volta a sviluppare meccanismi di spiegabilità. Dovrebbero anche essere disponibili spiegazioni sulla misura in cui un sistema di IA influenza e definisce il processo decisionale organizzativo, le scelte di progettazione del sistema e la logica alla base della sua diffusione, in modo da garantire la trasparenza non solo dei dati e dei sistemi, ma anche dei modelli di business. Infine, è importante comunicare, opportunamente e in modo adeguato al caso in esame, le capacità e i limiti del sistema di IA ai diversi portatori di interessi coinvolti. I sistemi di IA dovrebbero inoltre essere identificabili come tali, così che gli utenti sappiano che stanno interagendo con un sistema di IA e possano individuare le persone che ne sono responsabili*" Comunicazione "*Creare fiducia nell'intelligenza artificiale antropocentrica*" dell'8 aprile 2019 (COM(2019) 168 final).

[32] Nella proposta, l'idea di fondo è imporre regole stringenti, ed equivalenti per tutti gli Stati membri, proporzionate al tipo di sistema di I.A. e al rischio di lesione dei diritti fondamentali derivante dal suo utilizzo, o per il settore in cui si applica, o per le modalità e tecniche utilizzate. Vengono così individuati quattro livelli di rischio: 1.) *inaccettabile*, con divieto generale di applicazione (sistemi che manipolano il comportamento umano, con tecniche subliminali; sistemi volti ad attribuire un punteggio sociale da parte dei governi sui comportamenti e le manifestazioni del pensiero dei cittadini; sistemi che sfruttano la vulnerabilità di persone per l'età o la disabilità; pratiche biometriche di identificazione in luoghi aperti al pubblico; per queste ultime, tra cui rientrano i sistemi di riconoscimento facciale, l'applicazione è però consentita per esigenze di pubblica sicurezza indicate nel regolamento, previa autorizzazione rilasciata dall'autorità giudiziaria o da una autorità amministrativa indipendente); 2.) *rischio elevato,* cui è dedicato l'intero Titolo III, per i quali sono imposte regole vincolanti in termini di validazione, etichettatura, analisi dei rischi e la cui violazione può comportare ingenti sanzioni pecuniarie (i sistemi ad alto rischio, che costituiscono il fulcro centrale della disciplina, riguardano trasversalmente quasi tutti i settori di possibile impiego: trasporti, gestione di sistemi a rete, componentistica di sicurezza dei prodotti, selezione e valutazione del personale, erogazione dei servizi pubblici, ad esempio per l'accesso al credito, istruzione e formazione, attività di polizia di prevenzione o contrasto del crimine, sistemi di assistenza nell'amministrazione della giustizia); 3.) *rischio limitato*, come i *chatbot,* gli assistenti virtuali; 4.) *rischio minimo*, come i videogiochi, per i quali il regolamento non si applica, ma viene incentivata l'adozione di "codici deontologici" a garanzia del rispetto di standard minimi di affidabilità. Il modello della ibridazione tra regole giuridiche e regole tecniche, mediante

Confermando la natura complessa del principio, quale sintesi tra "conoscenza" e "comprensione", la proposta prevede che per tutti i sistemi di I.A -e quindi anche per quelli a rischio minimo come i *chatbot*[33] - è necessario garantire quantomeno la "conoscibilità" come sopra indicata; è cioè necessario assicurare che gli utenti siano resi edotti del fatto che stanno interloquendo con una macchina, in modo da consentire loro di decidere consapevolmente se continuare ad interagire oppure no. Questa garanzia minima conosce però due eccezioni: essa non si applica se ciò è ovvio in base alle circostanze o al contesto e se il sistema è autorizzato dalla legge con la finalità di prevenire, effettuare indagini o reprimere attività criminali, (ivi compresa l'ipotesi in cui vengano usati sistemi di riconoscimento biometrico o emozionali)[34].

Del principio di "esplicabilità" quale "comprensibilità" da parte dei destinatari dei risultati di I.A., si occupa invece l'art. 13, rubricato "*trasparenza e fornitura di informazioni agli utenti*", applicabile solo ai sistemi ad alto rischio. Per essi, deve essere garantito un livello di trasparenza "*sufficiente*" a "*consentire agli utenti di interpretare i risultati del sistema e utilizzarli in modo appropriato*". Non si parla più di "*logica*" del trattamento automatizzato, come nella disciplina europea della protezione dei dati personali, ma di "*interpretabilità*", così guardando alla finalità cui la trasparenza rafforzata deve tendere dal lato dei destinatari ed affermando una valenza generale del principio, cui devono rispondere – almeno sul piano normativo e quindi ai fini della ammissibilità degli stessi - anche i sistemi di I.A. che non sono sorretti da "tecniche" od "approcci" simbolici come quelli di Deep Learning, per i quali è difficilmente enucleabile una "logica" di ragionamento analoga a quella "giuridica"[35].

---

l'adozione di un "codice" vincolante i progettisti, è già utilizzato, proprio con riguardo al principio di spiegabilità degli algoritmi, dalla USACM (Associazione statunitense della meccanica computazionale), che, partendo dalla crescente evidenza che "*alcuni algoritmi e analisi possono essere opachi, rendendo impossibile determinare quando i loro output possono essere distorti o errati prevede espressamente*" e che "*le decisioni prese dagli algoritmi predittivi possono essere opache a causa di molti fattori, tra cui tecnici (l'algoritmo potrebbe non prestarsi a una facile spiegazione), economici (il costo di fornire trasparenza può essere eccessivo, compreso il compromesso dei segreti commerciali) e sociali (rivelando l'input può violare le aspettative sulla privacy)*" ha anch'esso enucleato sette *Principles for Algorithmic Transparency and Accountability* cui conformare la progettazione, compreso quello della "spiegabilità" *https://www.acm.org/binaries/content/assets/publicpolicy/2017_joint_statement_algorithms.pdf,* visitato il 10 giugno 2021. E' invece un atto amministrativo generale il "codice" cui si riferisce 13-bis del D.Lgs. 7 marzo 2005, n. 82 *Codice dell'amministrazione digitale*, introdotto dal già citato D.L. 76/2020, cd. Decreto semplificazione 2020, secondo cui i sistemi informatici e servizi digitali delle pubbliche amministrazioni devono essere progettati, realizzati e sviluppati "*in coerenza con gli obiettivi dell'agenda digitale italiana ed europea e nel rispetto del codice di condotta tecnologica"* (comma 1), specificando che "*il codice di condotta tecnologica disciplina le modalità di progettazione, sviluppo e implementazione dei progetti, sistemi e servizi digitali delle amministrazioni pubbliche, nel rispetto del principio di non discriminazione, dei diritti e delle libertà fondamentali delle persone e della disciplina in materia di perimetro nazionale di sicurezza cibernetica*" (comma 2).

[33] I *chatbot* sono programmi software che impiegano intelligenza artificiale per simulare un agente o assistente intelligente in una conversazione con un utente umano.

[34] Art. 52, par. 1 *Transparency obligations for certain AI systems*

[35] L'obiettivo si traduce però in gravosi oneri imposti ai soggetti, pubblici o privati, che si avvalgono di sistemi di I.A. ad alto rischio. Analogamente a come avviene per i prodotti pericolosi, i sistemi, per poter essere validati ed etichettati, prima di essere utilizzati, devono essere corredati da "*istruzioni per l'uso in un formato digitale appropriato o altrimenti che includano informazioni concise, complete, corrette e chiare che siano pertinenti, accessibili e comprensibili per gli utenti*". Le informazioni devono in particolare riguardare: l'identità e i dati di contatto di chi si avvale del sistema, a prescindere dalla sua natura

La lettura del proposto articolo 13 mette in luce, peraltro, anche un altro profilo. In esso si prevede che i sistemi ad alto rischio debbano essere "progettati" in modo da essere adeguatamente trasparenti. A prescindere dal contenuto del principio di esplicabilità, viene consacrata, pertanto, anche per il più complesso ambito della I.A. "antropocentrica", una nuova collocazione delle regole rispetto al fenomeno da regolare, che già era evincibile dall'art. 25 del GDPR (cd. Privacy by Design), attualizzata come "Ethics by Design" (EdB). Si tratta di un approccio alla progettazione dei sistemi algoritmici, destinati ad essere veicolati in attività disciplinate da norme e regole, che mira all'inclusione sistematica di valori, principi, requisiti e procedure etici nelle fasi di progettazione e di sviluppo e che implica un rovesciamento di prospettiva. La regola giuridica si integra, trasformandosi in un requisito tecnico, nella elaborazione del modello algoritmico che viene conformato fin dalla sua progettazione dalla norma giuridica; la regola giuridica in sostanza "gioca in anticipo", si colloca in un momento anteriore anche alla sua possibile violazione, delinea il fenomeno dall'interno. Come efficacemente osservato, "*si potrebbe dire che una tecnologia, potenzialmente lesiva, viene privata della sua offensività per via tecnica. Il diritto governa questo movimento tecnico, ma fa un (mezzo) passo indietro rispetto alla tradizionale logica violazione/reazione*"[36].

Teoricamente ineccepibile, tale cambio di passo comporta però concretamente un ripensamento anche delle competenze tradizionali del giurista; ovvero la necessaria acquisizione della capacità di confrontarsi con competenze tecniche diverse, di lavorare in assetti multidisciplinari, non secondo logiche al ribasso di forzate negoziazioni, ma con la piena consapevolezza che debbano individuarsi sostrati comuni, logiche condivise, obiettivi unitari, al fine di cogliere effettivamente le opportunità in termini di efficienza dell'amministrazione che l'era digitale può apportare, nel perimetro delineato dal rispetto dei principi fondamentali dell'azione amministrativa[37]; in sostanza, una rivoluzione culturale. Forse, tra tutte, la sfida più ambiziosa che l'I.A. lancia al mondo del diritto.

pubblica o privata; le caratteristiche, capacità e limitazioni delle prestazioni del sistema, tra cui lo scopo previsto, il livello di accuratezza, robustezza e sicurezza informatica; qualsiasi circostanza nota o prevedibile che possa comportare rischi per la salute e la sicurezza o per i diritti fondamentali; le sue prestazioni rispetto alle persone o ai gruppi di persone su cui si intende utilizzare il sistema; se opportuno, ("when appropriate"), specificazioni relative ai dati di input o qualsiasi altra informazione rilevante in termini di addestramento, validazione e test del seti di dati utilizzati, tenendo conto dello scopo previsto del sistema di IA. Si tratta di istruzioni generali, concernenti l'architettura del sistema e il set di dati oggetto di elaborazioni, che, ad una prima lettura, non sembrano però di per sé garantire il principio di "comprensibilità" che invece riguarda la concreta decisione in cui confluisce l'elaborazione operata dal sistema di I.A.

[36] A. SANTOSUSSO, *Intelligenza artificiale e diritto. Perché le tecnologie di IA sono una grande opportunità per il diritto*, cit., pag 189. Portata alle estreme conseguenze, la compenetrazione tra oggetto da regolare e norma porta a delineare il cd. "diritto computazionale", come quando si immagina che le future macchine a guida autonoma possano essere progettate avendo già all'interno inserite le regole del codice della strada, in modo da rendere impossibile la sua violazione. Con immaginabili conseguenze, qualora la vita reale presenti circostanze non previste o imprevedibili, come è accaduto per l'attraversamento di una donna che portava un carrello della spesa nell'incidente di Tempe del 2018.

[37] Nature, *Meet the challenge of interdisciplinary science. Problems of modern society demand collaborative reserche*, 30 giugno 2016, https://www.nature.com/articles/534589b, visitato il 10 giugno 2021.