L'IRCOCERVO
PRIMA RIVISTA ELETTRONICA ITALIANA DI METODOLOGIA GIURIDICA - TEORIA GENERALE DEL DIRITTO E DOTTRINA DELLO STATO
FIRST ITALIAN DIGITAL JOURNAL OF LEGAL METHODOLOGY - GENERAL THEORY OF LAW AND DOCTRINE OF THE STATE

# Spiegare le pieghe.
# Alcune osservazioni su trasparenza e spiegabilità

Michele Martoni

*Università degli Studi di Urbino*

***Abstract***: **Explaining the Folds. Some Remarks on Transparency and Explainability**
Governing a phenomenon implies that one must first be able to understand and explain it, literally peel back the folds, bringing it to light to evaluate and estimate its impacts. Specifically, within a significantly extensive and articulated framework, a primary issue that needs to be addressed concerns the ability to "look between the folds," scrutinize opacity, and reflect on the concept of black-box in machine learning systems. System transparency is a prerequisite for human-machine coexistence. In this direction, some considerations are intended to be placed on the explainability of the decision as an extension and implementation of the transparency principle. Explainability is a very complicated and technical matter, but one that lends itself to a preliminary principled evaluation. It is the response, or has the potential to be, to a question of meaning, a legal claim, and becomes a method for ethically interrogating innovation.

***Keywords***: Artificial Intelligence, Machine Learning, Transparency, Explainability, Trustworthiness.



## 1. Premessa

Se dobbiamo chiederci quale sia la differenza fra uomo e macchina, allora non abbiamo capito cos'è l'uomo oppure non abbiamo capito cos'è la macchina. Chi crede che "tra cervello umano e computer non possa esserci alcuna distinzione categoriale intacca i fondamenti non solo della prassi scientifica, ma in generale del mondo della vita"[1]. Come osserva Lombardi Vallauri:

[1] J. Nida-Rümelin, N. Weidenfeld, *Umanesimo digitale. Un'etica per l'epoca dell'Intelligenza Artificiale*, trad. it., FrancoAngeli, Milano, 2019, p. 27. Sulle qualità distinte dell'intelligenza umana e dell'intelligenza artificiale, vedi E.J. Larson, *Il mito dell'intelligenza artificiale. Perché i computer non possono pensare come noi*, trad. it., FrancoAngeli, Milano, 2022.

 ISSN 1722-392X
Special issue: M.N. Campagnoli, P. Marra (eds.),
*Artificial Intelligence and Neuro-cognitive Sciences in Law: From Symbiosis to Substitution*,
https://doi.org/10.57581/LIRCOCERVO.23.2024.1.13

> capire dei significati intelligibili è completamente diverso dal captare dei messaggi, come fanno invece i computer. I computer riconoscono dei significanti, ma non capiscono dei significati. La differenza tra i significanti e i significati è che i significanti sono dei pezzi di materia, mentre i significati sono del tutto al di fuori della materia[2].

Secondo Finocchiaro ad indurci in questo equivoco è una narrazione delle tecnologie che parla di intelligenza, di oracoli, di allucinazioni, di incantesimi, invece che di sistemi, di esiti delle ricerche, di errori, difetti o malfunzionamenti[3], vale a dire una narrazione dell'artificiale "intelligente" che presenta una serie di forzature, lessicali e sostanziali. Una narrazione che, peraltro, ci espone al rischio di non riuscire a maneggiare con la cura necessaria una materia così complessa, anche a fronte del profilarsi di possibili nuove soggettività giuridiche[4].

Per certo, è possibile mimare (agendo come) l'umano, aggirando o eludendo le peculiarità che fanno dell'umano ciò che è. Un esempio è rappresentato dai sistemi di traduzione[5]. Si elude l'umano, agendo nel mondo, al di fuori del mondo, creando esperienze al di fuori dell'esperienza.

La vera questione allora non sta nella differenza, per così dire ontologica tra uomo e macchina, bensì nel ruolo che si vuole far giocare all'agente macchinico (anche) in relazione all'agente umano. Non si tratta di sostanza, ma di governo, di scelte e di ruoli da assegnare[6]. La questione è di individuare, dunque, limiti e forme, *guard-rail* etici[7], confini[8], soglie[9] da non oltrepassare a protezione dell'*altro* (altro

[2] L. Lombardi Vallauri, "Algoretica e Informatica giuridica", in *I-lex – Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale*, 15 (2022), n. 1, p. 29. Si veda anche Id., "Algoretica. Le due sfide cruciali nell'era tecnologica: bioetica, roboetica", in *Atti e memorie dell'Accademia toscana di scienze e lettere La Colombaria*, LXVIII, 2017, pp. 355-376.

[3] G. Finocchiaro, *Intelligenza artificiale. Quali regole?*, il Mulino, Bologna, 2024, p. 25.

[4] *Ivi*, pp. 21 e ss. Si vedano anche G. Taddei Elmi, "I diritti dell''intelligenza artificiale' tra soggettività e valore: fantadiritto o *ius condendum*", in *Il Meritevole di tutela (Studi per una ricerca coordinata da Luigi Lombardi Vallauri)*, Giuffrè, Milano, 1990, pp. 685-711; Id., "L'intelligenza artificiale tra soggettività e valore: capacità cognitiva e capacità giuridica dei sistemi intelligenti", in A.A. Martino (a cura di), *I Sistemi Esperti Giuridici*, Cedam, Padova, 1989, pp. 916-944.

[5] G. Sartor, *L'intelligenza artificiale e il diritto*, Giappichelli, Torino, 2022, p. 22.

[6] Si vedano, fra gli altri, B. Romano, *Algoritmi al potere. Calcolo giudizio pensiero*, Giappichelli, Torino, 2018; S. Pozzolo, F. Cabitza, A. Rossetti (a cura di), "Governare l'intelligenza artificiale", in *Ragion pratica*, (2021), n. 2; P. Marra, "'Oltre la misura'. Considerazioni brevi su ragionamento e calcolabilità giuridica", in *Annali del Dipartimento Jonico*, (2022), pp. 113-128.

[7] P. Benanti (2021), *Audizione della commissione straordinaria per il contrasto dei fenomeni di intolleranza, razzismo, antisemitismo e istigazione all'odio e alla violenza*. Recuperato da https://www.senato.it/application/xmanager/projects/leg18/attachments/documento_evento_procedura_commissione/files/000/401/901/Audizione_prof._Benanti.pdf, per i resoconti: https://www.senato.it/leg18//1122?indagine=1601, [Data di consultazione: 30/05/2024]. Del medesimo autore si veda anche P. Benanti, *La condizione tecno-umana*, EDB, Bologna, 2022.

[8] T. Casadei, S. Pietropaoli, "Intelligenza artificiale: fine o confine del diritto?", in Ead. (a cura di), *Diritto e tecnologie informatiche*, Wolters Kluwer, Milano, 2021.

[9] M.P. Mittica, *Il pensiero che sente. Pratiche di Law and Humanities*, Giappichelli, Torino, 2022.

umano, altro animale non umano, altro in generale come ambiente naturale), per i quali serve un riferimento valoriale, un progetto umano[10].

Preliminarmente, tuttavia, è necessario comprendere e spiegare il problema, "spiegandone" letteralmente "le pieghe", per portarlo alla luce, valutarlo e stimarne le implicazioni. Per questo è essenziale chiarire che, seppure l'intelligenza artificiale non è identificabile con un cervello umano, allo stesso tempo non è riducibile a semplice strumento, a pura materia (un'interpretazione meccanicistica sarebbe parziale, dunque errata).

In tal senso pare appropriata la sintesi riferita dall'art. 3, paragrafo 1, del Regolamento UE sull'intelligenza artificiale, noto come *AI Act*[11] che definisce un sistema di intelligenza artificiale come "un sistema automatizzato", con "livelli di autonomia variabili", che può presentare capacità di adattarsi e che, "per obiettivi espliciti o impliciti", dall'*input* che riceve deduce "come generare *output* quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali".

Altrettanto utile è il ricorso alla nota categoria di Marcel Mauss, per cui l'intelligenza artificiale è osservata come un "fatto sociale totale", per la sua capacità di influenzare e determinare così tanti fenomeni da coinvolgere la gran parte dei meccanismi di funzionamento di tutti gli ambienti di vita e il loro mutamento[12].

Soltanto da una corretta individuazione del problema e del suo impatto è possibile, in altre parole, procedere con la indispensabile regolazione della costruzione (etica del programmatore) e delle possibilità di impiego (etica del programma), così come diviene imprescindibile elaborare le strategie utili per educare alla cittadinanza anche digitale[13].

[10] L. Floridi, *Il verde e il blu*, Raffaello Cortina Editore, Milano, 2020. Si veda anche Id., *Etica dell'intelligenza artificiale. Sviluppi opportunità, sfide*, Raffaello Cortina Editore, Milano, 2022.

[11] L'*Artificial Intelligence Act* (noto come *AI Act* o *AIA*) è il regolamento comunitario di prossima pubblicazione "che stabilisce regole armonizzate sull'intelligenza artificiale". È stato approvato il 13 marzo 2024 dal Parlamento europeo dopo un lungo *iter* avviato con la proposta di regolamento presentata dalla Commissione europea il 21 aprile 2021. La proposta è stata a sua volta oggetto di un lungo negoziato terminato con l'accordo fra il Consiglio e il Parlamento raggiunto nel mese di dicembre 2023. La pubblicazione del regolamento sulla Gazzetta Ufficiale dell'Unione europea dovrebbe avvenire nell'estate del 2024 dopo il via libera finale dal Consiglio. L'entrata in vigore è stabilita dall'art. 113 nel ventesimo giorno successivo alla pubblicazione. Il regolamento diventerà efficace 24 mesi dopo la data di entrata in vigore, salvo alcune eccezioni previste dall'art. 113.

[12] Cfr. M. Mauss, *Saggio sul dono*, trad. it., Einaudi, Torino, 2002. L'accostamento tra la trasformazione digitale e la categoria maussiana è di A. Garapon, L. Lassegue, *La giustizia digitale*, trad. it., il Mulino, Bologna, 2021, p. 79.

[13] Sulla educazione alla cittadinanza digitale si vedano, fra gli altri, A.C. Amato Mangiameli, M.N. Campagnoli, *Strategie digitali. #diritto_educazione_tecnologie*, Giappichelli, Torino, 2020; G. Pascuzzi, *La cittadinanza digitale. Competenze, diritti e regole per vivere in rete*, il Mulino, Bologna, 2021; V. Marzocco, S. Zullo, T. Casadei, *La didattica del diritto. Metodi, strumenti e prospettive*, Pacini giuridica, Pisa, 2021; B.G. Bello, *(In)giustizie digitali. Un itinerario su tecnologie e diritti*, Pacini giuridica, Pisa, 2023.

Va chiarito, infine, che la definizione di intelligenza artificiale non individua un'unica tecnologia. Si tratta di una galassia di diverse tecniche, con caratteristiche e applicazioni che variano.

Nel presente lavoro ci si riferirà soltanto al *machine learning*, cioè un sistema di apprendimento automatico basato sull'utilizzo dei dati. Essendo la trasparenza dei sistemi un presupposto necessario per la convivenza fra umano e macchinico, in un quadro decisamente esteso e articolato, una prima questione che necessita di essere trattata riguarda la possibilità, "guardando fra le pieghe", di scrutare nell'opacità della *black-box* dei sistemi di apprendimento automatico.

In questa direzione si intendono porre alcune considerazioni sulla "spiegabilità" della decisione quale estensione e attuazione del principio di trasparenza. La spiegabilità è questione molto complicata e tecnica, ma che si presta a una valutazione di principio preliminare: è la risposta, o ha le potenzialità per esserlo, a una domanda di senso portatrice di una pretesa giuridica, e può divenire metodo per interrogare eticamente l'innovazione.

Quanto alla scelta del termine "spiegabilità" va precisato che, a fronte delle definizioni di AI esplicabile, AI interpretabile e AI spiegabile[14], si è optato per il termine "spiegabilità" con riferimento alla "esplicabilità", ovvero alla "*explicability*", seguendo, da un lato, l'impostazione prescelta nella traduzione ufficiale degli *Orientamenti etici per un'IA affidabile*, redatti dal Gruppo indipendente di esperti ad alto livello sull'intelligenza artificiale istituito dalla Commissione europea nel giugno 2018[15], e dall'altro lato, la ragione suggerita dalla maggiore ampiezza dell'impiego del termine "spiegabilità" per indicare la cosiddetta *explainable AI* (XAI)[16].

A proposito di esplicabilità, inoltre, Floridi individua due accezioni di senso che risultano estremamente preziose per la nostra riflessione e faranno da corollario anche al nostro impiego della "spiegabilità":

> il senso *epistemologico* di *intelligibilità* (come risposta alla domanda: 'Come funziona?') [e] quello *etico* di responsabilità (*accountability*) (come risposta alla domanda: 'Chi è responsabile del modo in cui funziona?')[17].

[14] Sulla distinzione fra *explicability* e *interpretability*, si veda S. Ali, T. Abuhmed, S. El-Sappagh *et al*., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence", in *Information Fusion*, 99 (2023), p. 2. Su spiegabilità e esplicabilità, si veda L. Floridi, *Etica dell'intelligenza artificiale*, cit., pp. 91 e ss.

[15] Si v. per maggiori dettagli https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, [Data di consultazione: 30/05/2024].

[16] Si veda S. Ali, T. Abuhmed, S. El-Sappagh *et al.*, *op. cit*.

[17] L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 92. Si vedano anche L. Floridi, J. Cowls, "A unified framework of five principles for AI in society", in *Harvard Data Science Review*, (2019), n. 1, pp. 1-14; L. Floridi, J. Cowls, M. Beltrametti *et al*., "AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", in *Minds and Machines*, 28 (2018), pp. 689-707.

## 2. Fra le pieghe

Cominciamo con l'intendere le nozioni di *black-box* e opacità rispetto ai sistemi di intelligenza artificiale.

Come osserva Sartor fino a pochi decenni fa, si presumeva che per creare un sistema intelligente fosse essenziale fornire a priori tutte le informazioni necessarie affinché la macchina potesse affrontare i compiti assegnati. Si riteneva che queste informazioni dovessero essere rappresentate mediante linguaggi formali ed elaborate attraverso processi di ragionamento automatico[18].

Negli ultimi anni, l'interesse degli studiosi si è spostato verso le tecniche di apprendimento automatico (*machine learning*), dando origine a sistemi capaci di generare autonomamente la conoscenza (il modello) su cui basano le loro prestazioni.

Nei sistemi di *machine learning*, la conoscenza è costruita sulla base di esempi, informazioni e dati, che gli sono forniti o a cui lo stesso può attingere.

Si distingue un algoritmo discente che apprende usando gli esempi e costruisce il modello, e un algoritmo appreso o modello appreso che è il risultato dell'apprendimento.

Il modello predisposto dall'algoritmo discente è poi usato dall'algoritmo predittore, per fornire risposte che vengono usualmente indicate come predizioni[19].

L'apprendimento automatico svolge essenzialmente due compiti, la classificazione e la regressione, a seconda della natura della variabile di uscita (*output*).

Si potrebbe rappresentare il processo di apprendimento come una sorta di tunnel con variabili che entrano (*input*) e altre variabili che escono (*output*).

In ingresso ci sono esempi, informazioni e dati, che vengono forniti al sistema e di cui si è già detto sopra.

L'uscita è costituita dalla variabile che si vuole "predire" che può essere una classe (nella classificazione), per esempio determinare se un'e-mail è *spam* o *non spam*, o un numero (nella regressione), per esempio predire il prezzo di una casa in base a diverse caratteristiche.

L'algoritmo giunge alla previsione con maggiore o minore accuratezza, senza necessariamente essere in grado di fornire una spiegazione dettagliata sul perché ha ottenuto quel risultato. Questo è particolarmente vero nel caso dei modelli più complessi.

Entra così in gioco il concetto di spiegabilità, di cui si dirà meglio successivamente, che negli ultimi anni ha visto un interesse crescente della ricerca scientifica principalmente di stampo tecnico.

Spiegare l'*output* significa individuare quali grandezze in ingresso o *features* (dati in *input*) hanno inciso maggiormente sulla decisione presa dall'algoritmo.

[18] G. Sartor, *op. cit.*, p. 35.
[19] *Ivi*, pp. 48 e 49.

Questo non è banale perché le *features* possono essere moltissime e in certi casi non è semplice capire, interpretare, conoscere il percorso decisionale seguito dall'algoritmo.

Gli algoritmi di classificazione (fra i quali, per esempio: gli alberi di decisione, le *support vector machine*, *naive bayes*, *random forest* e le reti neurali che nella loro accezione più complessa costituiscono il c.d. *deep learning*) non si comportano tutti nello stesso modo dal punto di vista della spiegabilità: alcuni sono più trasparenti mentre altri sono più opachi.

La complessità si accompagna alla opacità, più è complesso l'algoritmo, più è complicato spiegarne il funzionamento. D'altra parte maggiore è la complessità, più elevate sono le prestazioni che possiamo ottenere.

Vale la pena, pertanto, guardare più da vicino in cosa consistono questi algoritmi di classificazione e nello specifico gli alberi di decisione e le reti neurali.

L'albero di decisione è un classificatore che usa una struttura gerarchica ad albero, costituita da nodi di decisione e nodi foglia. Ciascun nodo di decisione rappresenta una scelta basata su una caratteristica o variabile di *input*, mentre i nodi foglia rappresentano le previsioni o le classificazioni finali. L'albero di decisione prende in considerazione una *feature* alla volta e, a seconda del valore della variabile, segue un certo percorso.

L'esempio che segue riguarda la classificazione di un messaggio come *spam*[20]:

```
(1) Lunghezza del messaggio <= 20?
|--- (True): Contiene la parola "Offerta"?
|    |--- (True): Spam
|    |--- (False): Non spam
|
|--- (False): Contiene la parola "Urgente"?
     |--- (True): Spam
     |--- (False): Non spam
```

L'albero si percorre dall'alto verso il basso. La prima decisione riguarda la lunghezza del messaggio. Se è minore o uguale a 20 caratteri, si passa alla seconda decisione basata sulla presenza della parola "Offerta". Se la lunghezza del messaggio è superiore a 20 caratteri, si passa alla terza decisione basata sulla presenza della parola "Urgente".

Le *features* in questo caso sono la lunghezza del messaggio, la presenza della parola "Offerta" e la presenza della parola "Urgente".

Questo algoritmo è trasparente e spiegabile dal momento che è possibile comprendere per quale ragione è stata presa la decisione specifica. Il processo decisionale rimane trasparente e spiegabile anche nel caso in cui vi sia un albero più articolato che può essere utile per una maggiore accuratezza.

[20] L'esempio riportato nel testo è stato ottenuto interrogando ChatGPT 3.5 con il seguente *prompt*: "puoi mostrarmi un esempio di albero di decisione?".

Le reti neurali sono algoritmi più complessi. Si tratta di un insieme di neuroni artificiali disposti in strati. Ogni neurone è connesso ai neuroni dei livelli precedenti e successivi. Dai neuroni del livello precedente riceve gli *input* che processa per generare l'*output*. Questa attività è regolata dai pesi associati per ogni neurone alle sue connessioni con i neuroni nei livelli precedenti, i quali vengono ottimizzati durante il processo di addestramento della rete per migliorarne le prestazioni.

Questo approccio permette alla rete neurale di individuare relazioni complesse senza la necessità di regole esplicite definite dall'utente come invece accade nel caso dell'albero decisionale.

Per quanto il processo decisionale interno alla rete possa essere più intricato rispetto all'albero decisionale, è possibile ricorrere a tecniche di spiegazione dei modelli (ad esempio visualizzando i pesi delle connessioni o analizzando l'importanza delle caratteristiche) per comprendere meglio come la rete neurale prende le proprie decisioni. Ciò nonostante, come osserva Sartor, questa informazione "non rende esplicite le ragioni per le quali è stata data una certa risposta" e non produce una "giustificazione comprensibile alla mente umana"[21]. Ed è ancora più vero nel caso del *deep learning*, ove vengono impiegate reti neurali artificiali più profonde e intricate, composte da numerosi strati nascosti e ciascuno con diversi neuroni.

Il *deep learning* è in grado di apprendere automaticamente rappresentazioni avanzate dei dati senza richiedere la definizione manuale delle regole o delle caratteristiche rilevanti come invece avviene negli alberi decisionali.

Si dovrebbe intendere, a questo punto, la ragione per cui le reti neurali sono dette "opache" o anche "scatole nere" (*black-box*), e in modo più esteso l'impiego in informatica delle espressioni *black-box*, *gray-box*, e *white-box*, per riferirsi ai diversi livelli di opacità dei componenti interni di un sistema.

L'opacità si associa alla complessità che a sua volta è collegata alle prestazioni. Vale a dire che maggiore è la complessità, più elevate sono le prestazioni e minore è la trasparenza del processo selettivo/decisionale del sistema.

Se così è, diventa allora evidente che in applicazioni e in contesti più delicati che impattano maggiormente sulla vita e i diritti dei soggetti è necessario andare oltre le prestazioni elevate, non essendo possibile accontentarsi di *output* di cui non si conosca la spiegazione.

Occorre allora riflettere su come procedere per rendere affidabile il sistema di intelligenza artificiale.

[21] G. Sartor, *op. cit.*, p. 59.

## 3. Procedere per principî. La trasparenza "del sistema"

Come osserva Laura Palazzani, la tecnologia non è un destino[22]. Il suo *design* è nelle mani dell'uomo.

Certamente la consapevolezza che tutto è in rapido divenire e opaco, provoca un senso di smarrimento.

Serve allora una bussola ben bilanciata per recuperare la direzione e, con quella, una certa fiducia.

Procedere privilegiando i principî potrebbe essere la "bussola" adatta perché consentirebbe al diritto di recuperare il suo ruolo profetico e creativo, in un certo senso "invertendo il passo" rispetto alla corsa affannosa, di cui siamo tutti testimoni, all'inseguimento di un'innovazione tecnologica da regolamentare rispetto alla quale è sempre in ritardo. Ha ragione Pagallo quando suggerisce di utilizzare i principî come una lente concettuale per pensare l'etica applicata alla società algoritmica[23].

Occorre allora domandarsi da quali principî procedere[24].

Per uscire dal "gioco dell'imitazione" ideato da Turing[25] e per svelare il trucco della stanza cinese immaginato da Searle[26], il primo passo da compiere è quello di "far luce" su queste strategie e la scelta di avviare la riflessione dal principio della trasparenza, fuor di metafora, si è rivelata quella più opportuna. Se ne ha conferma, peraltro, dal gruppo indipendente di esperti istituito dalla Commissione europea nel giugno 2018[27] che, nel definire gli orientamenti etici per

[22] Così nel suo intervento introduttivo in occasione del seminario tenuto da Antonio Punzi il 18 marzo 2024 su *L'esperienza giuridica e il nuovo ordine delle intelligenze*, nell'ambito del Ciclo di seminari 2023/2024 su intelligenza artificiale e diritto, presso Lumsa. Si veda anche L. Palazzani, *Tecnologie dell'informazione e intelligenza artificiale. Sfide etiche al diritto*, Studium, Roma, 2020.
[23] U. Pagallo, "Algoritmi e conoscibilità", in *Rivista di filosofia del diritto*, 9 (2020), n. 1, p. 102. Si veda anche L. Floridi, J. Cowls *et. al*., "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", in *Minds and Machine*, 28 (2018), n. 4, pp. 689-707.
[24] Per un quadro unificato dei principî etici per l'IA si veda L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 91.
[25] Alan Turing si serve del cosiddetto "gioco dell'imitazione" per rispondere alla domanda "possono pensare le macchine". Scopo del gioco per l'interrogante – chiuso in una stanza – è quello di determinare quale delle altre due persone al di fuori della stanza, sia l'uomo e quale la donna. Turing immagina poi che una macchina prenda il posto dell'uomo. La questione sarà allora se sia possibile per l'interrogante capire quale sia la donna e quale la macchina. In A.M. Turing, *Intelligenza meccanica*, trad. it., Bollati Boringhieri, Torino, 1994, pp. 121 ss.
[26] J.R. Searle, "La mente è un programma?", in *Le Scienze*, (1990), n. 259, pp. 16-21. Si veda anche G. Lolli in A.M. Turing, *op. cit.*, pp. 18 e 19.
[27] Si vedano *Orientamenti etici per un'IA affidabile*, documento redatto dal Gruppo indipendente di esperti ad alto livello sull'intelligenza artificiale istituito dalla Commissione europea nel giugno 2018. Per maggiori dettagli https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, [Data di consultazione: 30/05/2024].

un'intelligenza artificiale affidabile, ha individuato fra i requisiti o principî fondamentali la trasparenza e la spiegabilità[28].

La trasparenza riguarda i dati, i modelli di *business*, il sistema nel suo insieme. Essa è intimamente connessa alla tracciabilità, alla comunicazione e, come vedremo, alla spiegabilità.

La tracciabilità riguarda i *set* di dati e i processi che determinano la decisione, nonché gli algoritmi utilizzati. Tutto dovrebbe essere documentato secondo i migliori *standard* per consentire la tracciabilità e aumentare la trasparenza.

La comunicazione riguarda la trasparenza dei sistemi di intelligenza artificiale rispetto alla loro natura e la loro identificabilità in quanto artificiali. Ciò comporta che i sistemi di intelligenza artificiale non possano presentarsi agli utenti come esseri umani e che gli esseri umani hanno il diritto di sapere che stanno interagendo con sistemi artificiali.

Questi orientamenti sono stati recepiti nel testo del già richiamato *AI Act* che considera la trasparenza come un principio ed un requisito essenziale.

In particolare, il Capo IV è dedicato agli obblighi di trasparenza per i fornitori e i *deployer*. L'art. 50 al paragrafo 1 dispone che i fornitori garantiscano che i sistemi di intelligenza artificiale "destinati a interagire direttamente con le persone fisiche" siano progettati e sviluppati in modo tale che le persone fisiche interessate siano informate del fatto di stare interagendo con un sistema di intelligenza artificiale, a meno che ciò non risulti evidente dal punto di vista di una persona fisica ragionevolmente informata, attenta e avveduta, tenendo conto delle circostanze e del contesto di utilizzo.

Il successivo paragrafo 2 dispone che i fornitori di sistemi di intelligenza artificiale "che generano contenuti audio, immagine, video o testuali sintetici", devono garantire che gli *output* siano "marcati in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente".

Il paragrafo 4 impone ai *deployer* di un sistema di intelligenza artificiale che genera o manipola immagini o contenuti audio o video che costituiscono un "deep fake" di rendere noto che il contenuto è stato generato o manipolato artificialmente.

Del pari, il successivo paragrafo 5 obbliga il *deployer* di un sistema di intelligenza artificiale che genera o manipola testo pubblicato allo scopo di informare il pubblico su questioni di interesse pubblico, a rendere noto che il testo è stato generato o manipolato artificialmente. Tale obbligo non trova applicazione nel caso in cui il contenuto generato dall'intelligenza artificiale sia stato sottoposto a un processo di revisione umana o di controllo editoriale e vi sia un soggetto che è responsabile editoriale della pubblicazione del contenuto.

Il regolamento suggerisce, poi, l'elaborazione di codici di buone pratiche a livello dell'Unione per facilitare l'efficace attuazione degli obblighi relativi alla rilevazione e all'etichettatura dei contenuti generati o manipolati artificialmente.

[28] Si veda anche L. Floridi, *Etica dell'intelligenza artificiale*, cit., pp. 91 ss.

Infine, anche l'art. 13, rubricato "Trasparenza e fornitura di informazioni ai Deployer", dispone che i sistemi di intelligenza artificiale ad alto rischio siano progettati e sviluppati in modo tale da garantire che il loro funzionamento sia sufficientemente trasparente così da consentire ai *deployer* di interpretare l'*output* del sistema e utilizzarlo adeguatamente.

Una lettura attenta dell'*AI Act* richiede un'analisi in combinato disposto con il regolamento (UE) n. 679 del 2016 (noto come GDPR)[29] in quanto i sistemi di *machine learning* si alimentano di dati molto spesso personali. Anche il GDPR pone il principio di trasparenza fra i principî fondamentali incorporati nell'art. 5. Pizzetti osserva come il principio di trasparenza, in particolare nell'ambito del digitale, sia vicino al principio di lealtà[30]. La trasparenza si potrebbe allora intendere non solo come semplice apertura, ma come qualcosa che aprendosi si vuol far comprendere, anche rispetto ai propri effetti.

Come scrive Califano, la "trasparenza, che rappresenta senz'altro un presupposto fondamentale dell'autodeterminazione informativa […], riguarda [poi] l'intera architettura dell'*accountability*" e dunque si connette al modello *risk based* sul quale poggia il principio di responsabilizzazione[31]. Tale impostazione si ritrova nell'art. 24 del GDPR che impone al titolare del trattamento di mettere in atto misure tecniche ed organizzative adeguate rispetto alla probabilità e gravità dei rischi per diritti e libertà degli interessati. Completano il quadro, e chiariscono la interazione con la trasparenza dei sistemi di AI, i principî della *privacy by design* (fin dalla progettazione) e della *privacy by default* (per impostazione predefinita) previsti dall'art. 25 del GDPR che hanno come orizzonte il rischio connesso al trattamento.

L'economia del presente contributo non consente purtroppo di dar conto anche della complessa trama normativa che si sta delineando per effetto di ulteriori interventi del legislatore comunitario (fra gli altri, per esempio, il *Digital Service Act*) caratterizzati da un approccio basato sul rischio e dall'*accountability* quali presupposti per l'effettiva tutela di diritti e libertà individuali. Certo è che, la trasparenza, che si pone come condizione preliminare di tutela, rischia di rivelare, come fa notare Han, una natura paradossale. Nel contesto contemporaneo che egli connota come "epoca dell'infocrazia", Han osserva la trasparenza come un dispositivo del potere, per cui quanto più si realizza la trasparenza, tanto più si dà spazio al dominio. Proprio la trasparenza sarebbe la causa dell'asimmetria informativa fra l'utente che è trasparente al controllo, e la tecnologia che vi si

[29] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE, noto come Regolamento generale sulla protezione dei dati o GDPR.

[30] F. Pizzetti, *Privacy e il diritto europeo alla protezione dei dati personali*, I, Giappichelli, Torino, 2016, p. 268.

[31] L. Califano, V. Fiorillo, F. Galli, *La protezione dei dati personali: natura, garanzie e bilanciamento di un diritto fondamentale*, Giappichelli, Torino, 2023, p. 83.

sottrae rimanendo chiusa e opaca[32]. Si potrebbe indicare questa accezione come il lato oscuro della trasparenza. "La prigione digitale è trasparente"[33], nel senso che, divenendo integralmente trasparente, il soggetto è dominato. Rodotà ne scriveva in termini di sindrome impiegando la metafora del *pesce rosso*. Il dominio, al contrario, non è trasparente: è un processo che si sottrae alla visibilità. Per Han: "la sala operativa della trasparenza è oscura", è una "*black box* algoritmica"[34].

Contestualmente, ulteriore paradosso, mentre rivela la sua oscura trasparenza, la tecnologia risponde a un bisogno di apertura, *disclosure* direbbero gli anglosassoni, dettato dalla sua progressiva e inarrestabile espansione. Così diviene sempre più *open*, guidando non per caso un nuovo "modo", costellato di *open source*, *open data* e *open government*, *open access*.

È da questa traiettoria che si intende trattare del principio di trasparenza, osservando come la tecnologia si apre, si mostra e inizia a raccontarsi, ma ancora non si spiega. Per questo sarà importante entrare nel merito della spiegabilità che, insieme alla tracciabilità e alla comunicazione, rimane strettamente connessa al principio di trasparenza.

## 4. La spiegabilità della decisione come presupposto di affidabilità

La spiegabilità attiene alla capacità di spiegare sia i processi tecnici di un sistema di intelligenza artificiale sia le relative decisioni. Come affermato negli Orientamenti etici per un'IA affidabile, "affinché un sistema possa essere tecnicamente spiegabile, gli esseri umani devono poter capire e tenere traccia delle decisioni prese dal sistema stesso" [35].

Per comprendere meglio l'ambito della spiegabilità è utile la guida predisposta dall'Information Commissioner's Office e The Alan Turing Institute intitolata *Explaining decisions made with AI* ove sono stati individuati sei tipi di spiegazione che si distinguono a loro volta tra spiegazioni basate sul processo (*process-based*) e spiegazioni basate sui risultati (*outcome-based*)[36]. I sei tipi di spiegazione individuati sono:

[32] B.-C. Han, *Infocrazia*, trad. it., Einaudi, Torino, 2023, p. 8.
[33] *Ivi*, p. 9.
[34] *Ivi*, p. 10.
[35] Si v. per maggiori dettagli https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, p. 20, [Data di consultazione: 30/05/2024].
[36] L'*Information Commissioner's Office* (ICO) e *The Alan Turing Institute* hanno redatto un documento intitolato *Explaining decisions made with AI* che vuole fornire alle organizzazioni consigli pratici per spiegare i processi, i servizi e le decisioni forniti o assistiti dall'IA alle persone interessate. Per maggiori dettagli si veda https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/, [Data di consultazione: 30/05/2024]. Si veda anche M. Van Lent, W. Fisher, M. Mancuso, "An Explainable Artificial Intelligence system for small-unit tactical behavior", in *Proceedings of the National Conference on Artificial Intelligence*, MIT Press, Cambridge (MA), 2004, pp. 900-907.

> *Rationale explanation*, ossia le ragioni che hanno condotto alla decisione; *Responsibility explanation*, chi è coinvolto nello sviluppo, nella gestione e nell'implementazione del sistema e chi contattare; *Data explanation*, ossia quali dati sono stati usati in una specifica decisione e come; *Fairness explanation*, ossia le misure adottate durante la progettazione e l'implementazione del sistema per garantire che le decisioni da esso supportate siano generalmente imparziali ed eque; *Safety and performance explanation*, ossia le fasi di progettazione e implementazione del sistema per massimizzare l'accuratezza, l'affidabilità, la sicurezza e la solidità delle sue decisioni e dei suoi comportamenti; e, infine, *Impact explanation*, ossia le misure adottate durante la progettazione e l'implementazione del sistema per considerare e monitorare l'impatto che l'uso dell'intelligenza artificiale e le sue decisioni hanno o possono avere su un individuo e su una società più ampia[37].

Con il tempo si è andato delineando un sempre più esteso ambito di ricerca noto come *Explainable AI* o XAI, che cerca di meglio comprendere questi meccanismi e indaga possibili soluzioni[38].

I metodi XAI come ognuno intende sono essenziali per la loro capacità di spiegare l'azione dei sistemi di intelligenza artificiale, e così facendo contribuiscono a realizzare soluzioni di intelligenza artificiale affidabili (in quanto verificabili nelle diverse componenti), anche in termini di *accountability*.

Questi metodi sono certamente utili per supportare gli esperti di dominio nei loro processi decisionali e di valutazione degli *output*, nonché per gli sviluppatori che, ove gli *output* si dovessero rivelare errati, sarebbero ulteriormente motivati ad indagare il funzionamento del sistema. Come abbiamo già detto, esiste una tensione fra la creazione di modelli spiegabili e la creazione di modelli con alte prestazioni. Il modello più accurato è, di norma, quello più complesso e con la minore spiegabilità.

Come si è già avuto modo di osservare, la spiegabilità è, dunque, questione complessa per la sua natura tecnica difficilmente riassumibile in poche righe,

[37] Si tratta di una mia traduzione letterale del testo https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/what-goes-into-an-explanation, p. 21.

[38] Per uno studio completo e una panoramica sulle ricerche e sulle tendenze in questo settore emergente si veda S. Ali, T. Abuhmed, S. El-Sappagh *et al.*, *op. cit.* Per un approccio informatico-giuridico e filosofico-giuridico si vedano, fra gli altri, M. Palmirani, S. Sapienza, "Big Data, Explanations and Knowability", in *Ragion pratica*, (2021), n. 2, pp. 349-364; M. Palmirani, "Big Data e conoscenza", in *Rivista di filosofia del diritto*, (2020), n. 1, pp. 73-92; U. Pagallo, "Algoritmi e conoscibilità", cit.; M. Durante, *Potere computazionale*, Meltemi, Milano, 2019. Si vedano anche G. Lo Sapio, "La black box: l'esplicabilità delle scelte algoritmiche quale garanzia di buona amministrazione", in *Federalismi.it*, 16 (2021), pp. 114-127; S. Dorigo (a cura di), *Il ragionamento giuridico nell'era dell'intelligenza artificiale*, Pacini giuridica, Pisa, 2020.

tuttavia, va accostata comunque, se non altro per una valutazione di principio preliminare.

Ciò che rende particolarmente interessante la spiegabilità, è la sua caratteristica di possedere tutte le potenzialità per essere la risposta a una domanda di senso, a una pretesa giuridica e, proprio per questa sua caratteristica, può rappresentare un valido metodo per interrogare eticamente l'innovazione.

Durante osserva che è in gioco l'idea stessa di libertà, in un contesto nel quale il decidere non è più un'esclusiva dell'essere umano. Infatti, nella misura in cui "non siamo in grado di spiegare la logica e il funzionamento, sentiamo di perdere contatto con la libertà" [39].

Pagallo scrive che affinché l'intelligenza artificiale sia benefica e non malefica, "bisogna preliminarmente capire e spiegare il bene, o il danno, che l'IA può provocare o promuovere nella società, oltre a cogliere i modi in cui l'IA si presta concretamente a fare del bene o del male". "Allo stesso modo, per stabilire se l'IA promuova o non intacchi l'autonomia umana" è necessario conoscere come l'intelligenza artificiale agirebbe. Infine affinché quest'ultima sia giusta c'è da garantire "che chi sviluppa e impiega questa tecnologia ne sia responsabile", anche per capire le ragioni di quanto accaduto.

In tal senso, conclude Pagallo

> tanto nel campo della responsabilità civile extra-contrattuale, quanto nel settore penale, [...] diventa sempre più urgente garantire alle parti lese, o agli imputati, la disponibilità di dati che spieghino come ha funzionato l'ecosistema algoritmico e che debba rispondere per come l'ecosistema ha funzionato[40].

In letteratura si è sviluppato un nutrito dibattito circa la possibile configurazione di un diritto alla spiegabilità rispetto alla applicazione dell'art. 22 del GDPR[41].

Stando alla lettera della norma, l'art. 22, paragrafo 1, del GDPR, stabilisce che l'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona.

[39] M. Durante, *op. cit.*, pp. 297-298.
[40] U. Pagallo, "Algoritmi e conoscibilità", cit., p. 105.
[41] Si vedano, per esempio, M. Durante, *op. cit.*, pp. 296 ss.; S. Wachter, B. Mittelstadt, L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", in *International Data Privacy Law*, 2017; B. Goodman, S. Flaxman, "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation'", in *AI Magazine*, 38 (2017), n. 3, pp. 50-57; U. Pagallo, "Algo-Rhythms and the Beat of the Legal Drum", in M. D'Agostino, M. Durante (eds.), "The Governance of Algorithms", in *Philosophy & Technology*, 31 (2018), n. 4, pp. 507-524; M.E. Kaminski, "The right to explanation, explained", in *Berkeley TechnologyLaw Journal*, 34 (2019), n. 1, pp. 189-218.

Al paragrafo 2 dello stesso articolo si dispone che il limite del paragrafo 1 non si applica nel caso in cui la decisione sia necessaria per la conclusione o l'esecuzione di un contratto tra l'interessato e un titolare del trattamento, se sia autorizzata dal diritto dell'Unione o dello Stato membro cui è soggetto il titolare del trattamento, o se si basi sul consenso esplicito dell'interessato.

Il successivo paragrafo 3 dispone che nei casi in cui non si applica il diritto a non essere sottoposto a trattamenti automatizzati, il titolare del trattamento deve attuare misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato, come (almeno) il diritto di ottenere l'intervento umano da parte del titolare del trattamento, di esprimere la propria opinione e di contestare la decisione.

In letteratura si è dibattuto sul fatto che la norma non preveda espressamente un diritto alla spiegabilità[42]. Tale riferimento è contenuto solo nel considerando n. 71 laddove si afferma che, in ogni caso, il trattamento "dovrebbe essere subordinato a garanzie adeguate, che dovrebbero comprendere la specifica informazione all'interessato e il diritto di ottenere l'intervento umano, di esprimere la propria opinione, di ottenere una spiegazione della decisione conseguita dopo tale valutazione e di contestare la decisione".

Come noto i considerando hanno natura non vincolante. Tuttavia, a sostegno dell'esistenza del diritto è giunta anche l'Opinione WP251rev.01 recante "Linee guida sul processo decisionale automatizzato relativo alle persone fisiche e sulla profilazione ai fini del regolamento 2016/679", adottate dal *Working Party art. 29* ove si afferma che "il titolare del trattamento dovrebbe trovare modi semplici per comunicare all'interessato la logica o i criteri sui quali si basa l'adozione della decisione". Inoltre, si osserva che il regolamento impone al titolare del trattamento di "fornire informazioni significative sulla logica utilizzata, ma non necessariamente una spiegazione complessa degli algoritmi utilizzati o la divulgazione dell'algoritmo completo". Le informazioni fornite dovrebbero tuttavia essere "sufficientemente complete affinché l'interessato possa comprendere i motivi alla base della decisione" [43].

L'esistenza di un diritto alla spiegazione sarebbe sostenuta infine dai principî generali di responsabilizzazione e di trasparenza, oltre che dalle norme che stabiliscono la protezione dei dati *by design* e *by default*, insieme alla valutazione preventiva d'impatto[44].

Sebbene, "la tutela contro i danni alle persone provocati dall'impiego di algoritmi non passa necessariamente attraverso la protezione dei loro dati personali", il modello del GDPR potrebbe comunque, secondo Pagallo, rappresentare un punto di partenza per affrontare la questione, facendo leva "su un insieme di norme già in vigore[45].

[42] Si veda, fra gli altri, S. Wachter, B. Mittelstadt, L. Floridi, *op. cit.*
[43] Per un commento approfondito si veda U. Pagallo, *op. cit.*, p. 96.
[44] *Ivi*, p. 98.
[45] *Ivi*, pp. 100 e 101.

Più critico appare Durante, che scorge possibili limiti alla spiegabilità nei diritti di proprietà intellettuale, segreti industriali o di stato, e nelle caratteristiche tecniche di alcuni processi decisionali algoritmici [46].

La presenza di questi ostacoli, se da un lato ha sollevato dubbi sull'attuazione del principio della spiegabilità, dall'altro lato costituisce un pungolo a ripensare l'idea stessa di spiegazione. Ed è su questa considerazione che ci si avvia ad alcune brevi note conclusive.

## 5. Brevi note conclusive

Alle tre ideate da Asimov, Frank Pasquale suggerisce di aggiungere una quarta legge della robotica, in base alla quale "un robot deve sempre indicare l'identità del suo creatore, controllore o proprietario" [47]. La spiegabilità resta fra le righe, come fatto squisitamente umano che richiede l'individuazione di una sorta di responsabilità *by design* in capo a un soggetto che possa rispondere dell'operato della macchina[48].

Diversamente interviene direttamente sulla spiegabilità il gruppo di lavoro del Berkman Klein Center *for Internet & Society* presso l'Università di Harvard[49]. L'idea è di delegare ai sistemi automatizzati di intelligenza artificiale non soltanto la decisione ma anche la sua spiegazione. In sostanza anche la spiegazione sarebbe automatizzata. Ciò implicherebbe, come osserva Durante, la necessità di chiarirsi sull'idea di spiegazione in termini tali da poterne implementare la nozione in un sistema automatizzato[50].

Quest'ultima linea di indagine ci porta dritti al punto da cui ripartire, ovvero alla distinzione tra spiegabilità e trasparenza. Una cosa, infatti, è spiegare una decisione, un'altra, del tutto diversa, è conoscere i processi che attraversano un sistema di intelligenza artificiale.

Distinguere spiegabilità e trasparenza potrebbe consentire di salvare la spiegabilità a scapito di parte della trasparenza quando, per esempio, quest'ultima, per ragioni giuridiche (diritti altrui) o tecniche non sia conseguibile del tutto.

[46] M. Durante, *op. cit.*, p. 304.

[47] F. Pasquale, "Toward fourth law of robotics: preserving attribution, responsibility, and explainability in an algorithmic society", in *Ohio State Law Journal*, 78 (2017), pp. 1-12. Si veda anche J. Balkin, "The Three Laws of Robotics in the Age of Big Data", in *Ohio State Law Journal*, 78 (2017), pp. 1-45.

[48] M. Durante, *op. cit.*, pp. 307 ss.

[49] Per un approfondimento della tematica si veda F. Doshi-Velez, M. Kortz, R. Budish *et al.*, "Accountability of AI Under the Law: The Role of Explanation", in *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper*, 2017. Recuperato da https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf, [Data di consultazione: 30/05/2024].

[50] M. Durante, *op. cit.*, p. 315.

Riaffermare un principio non significa arrestare l'innovazione, ma non significa nemmeno calarlo dall'alto immutato e immutabile. Comporta piuttosto assumersi l'onere di comprendere il contesto, prospettare il futuro e, se serve, immaginare nuovi modelli.

Solo chi non ha direttrici valoriali, le competenze necessarie e un pizzico di creatività, costruisce barriere, emana norme su norme, dissemina (più o meno scientemente) inciampi burocratici, o, all'opposto, lascia campo libero a tutto ciò che la tecnologia e la corsa al profitto immaginano di poter realizzare.

L'innovazione deve essere accompagnata. È, dunque, come già detto, una questione di ruoli, di limiti, di governo, di capacità di affermare un progetto umano con un impianto valoriale chiaro e possibilmente ampiamente condiviso.

 