AI Act, trasparenza ed *explainable AI*

Stefano Tramacere
Scuola Superiore Sant'Anna, Pisa. Mail: Stefano.Tramacere@santannapisa.it

Marzio Di Vece
Scuola Normale Superiore, Pisa. Mail: marzio.divece@sns.it

Clara Punzi
Scuola Normale Superiore, Pisa. Mail: clara.punzi@sns.it

Dino Pedreschi
Università di Pisa. Mail: dino.pedreschi@unipi.it

Fosca Giannotti
Scuola Normale Superiore, Pisa. Mail: fosca.giannotti@sns.it

Introduzione

A partire dal 2018, la Commissione Europea ha elaborato una strategia complessiva orientata a promuovere un'"*IA antropocentrica, sicura, affidabile ed etica*". Nel 2019, il Gruppo di Esperti sull'IA (*AI HLEG*) – organismo tecnico-consultivo istituito dalla Commissione – ha elaborato le *"Linee Guida Etiche per un'IA Affidabile"*[1], costituendo l'intelaiatura concettuale dell'approccio europeo alla regolamentazione dell'IA. Il risultato dello sforzo regolatorio è l'Artificial Intelligence (AI) Act[2] che pone la trasparenza come requisito fondamentale al fine di minimizzare i rischi per la salute, la sicurezza e salvaguardare i diritti fondamentali della persona. L'AI Act identifica quattro principi fondativi: il rispetto dell'autonomia umana, la prevenzione dei danni, l'equità e l'esplicabilità (o spiegabilità). Il principio dell'esplicabilità viene declinato con maggiore specificità all'interno del requisito etico della trasparenza. Quest'ultimo comprende tre dimensioni interconnesse: la tracciabilità, la spiegabilità e la comunicazione informativa e pertanto, si estende lungo l'intero ciclo di vita di un sistema di IA - dalla fase di progettazione e sviluppo fino all'implementazione operativa ed al monitoraggio nel tempo- e non si limita alla sola comprensibilità degli algoritmi. Secondo l'AI Act, la trasparenza riguarda quindi tutti gli elementi costitutivi del ciclo di vita di un sistema che possano facilitare la comprensione del suo funzionamento da parte dei soggetti interessati, come stabilito dai requisiti della sezione 2 del Capitolo III.

La trasparenza nell'IA: spiegabilità e interpretabilità. Sul lato della ricerca ed innovazione tecnico scientifica dell'Intelligenza Artificiale, sin dal 2015, si sviluppa una sotto-area di ricerca denominata XAI (eXplainable AI) per far fronte alla crescente complessità e opacità dei sistemi di *deep learning* accompagnata dalla fragilità di queste "scatole nere" di apprendere ed amplificare possibili stereotipi (*bias*) ereditati dai dati di allenamento. L'obiettivo della XAI è sviluppare tecniche che rendano comprensibile il

[1] European Commission and Directorate-General for Communications Networks, Content and Technology (2019) *Ethics guidelines for trustworthy AI.* Publications Office of the European Union https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[2] European Union, AI Act Regulation 2024/1689 – Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), Publications Office of the European Union, 2025, https://data.europa.eu/doi/10.2804/4225375.

processo decisionale che ha determinato il risultato finale dei modelli di machine learning. La XAI è il mattone fondamentale per supportare collaborazioni affidabili e sinergiche tra persona e macchina: cruciale per preservare l'autonomia della persona e metterla in controllo del "perché", "perché no" e "cosa potrebbe essere cambiato". Il presupposto è che l'IA possa costituire un valido supporto al processo decisionale dell'essere umano solo se quest'ultimo è posto nelle condizioni di intenderne adeguatamente il funzionamento e di riporre in essa una fiducia proporzionata, fondata cioè sulla corretta comprensione tanto delle potenzialità quanto dei limiti del sistema.

Secondo il paradigma dell'XAI, il suggerimento del modello di AI – una previsione, una raccomandazione o altro – prodotto in risposta ad una *richiesta*, è arricchito con un ulteriore elemento informativo, la *spiegazione,* che si concentra sul "*perché di quel suggerimento*". Questa spiegazione può assumere diverse forme: evidenziare quali elementi della richiesta hanno maggiormente influenzato il modello a restituire quel suggerimento, oppure la spiegazione può consistere in esempi, oppure evocare cosa si potrebbe modificare per ottenere un diverso suggerimento, supportando quindi una forma di ragionamento controfattuale. La forma della spiegazione ed il linguaggio in cui essa è espressa dipendono dal tipo di utente a cui essa è rivolta: gli sviluppatori che vogliono verificare il buon funzionamento del sistema, decisori esperti dei domini applicativi, talvolta scienziati dei dati, abituati a linguaggi scientifici, utenti finali dei servizi di quei domini applicativi che necessitano di interazioni in linguaggio naturale o metafore visuali, infine figure legali incaricate di verificare la conformità di quei servizi.

Le tecniche XAI studiate per produrre questo ulteriore elemento informativo si dividono in due categorie principali[3,4,5]:

Post-hoc: aggiungono un "*modulo di spiegazione*" ad un modello esistente e producono la spiegazione interagendo con il modello stesso al fine di ricostruire a posteriori la logica usata. La metafora di riferimento è "aprire le scatole nere/opache"

Interpretabili per disegno: creano modelli che sono intrinsecamente comprensibili, nel senso che la restituzione della spiegazione è estraibile direttamente da essi senza l'aggiunta di ulteriori componenti. La metafora di riferimento è "disegnare scatole bianche/trasparenti".

In entrambe le situazioni, esistono due momenti critici per costruire la spiegazione: il primo è l'estrazione della spiegazione dai modelli, una sorta di "artefatto" e poi la sua presentazione all'utente da confezionarsi nel linguaggio appropriato per il suo profilo.

L'efficacia dei metodi di XAI dipende dal contesto di applicazione, dal livello di rischio, dai diritti in gioco e dal destinatario delle informazioni. L'area non ha ancora raggiunto un livello di maturità tecnologica per tutte le tipologie di dato e di modello, ma esistono librerie che permettono agli sviluppatori di sistemi di AI di inserire la funzionalità di spiegazione, sperimentarne le

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, *A survey of methods for explaining black box models*. ACM computing surveys (CSUR), 51, 5, 2018, 1-42. http://dx.doi.org/10.1145/3236009.

[4] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, *Benchmarking and survey of explanation methods for black box models* in *Data Mining and Knowledge Discovery*, 37, 5, 2023, 1719-1778. https://doi.org/10.1007/s10618-023-00933-9

[5] C. Rudin, *stop explaining black box machine learning models for high stakes decisions and use interpretable models instead* in *Nature machine intelligence*, 1, 5, 2019, 206-215. https://doi.org/10.1038/s42256-019-0048-x.

diverse opzioni, studiare le misure di qualità che sono state sviluppate in questi anni, ed iniziano ad essere una componente solida in molti sistemi industriali.

Fino a qua abbiamo descritto il contesto di XAI nel caso di *Intelligenza Artificiale Discriminativa o, meglio,* di metodi di machine learning che imparano a discriminare tra diverse opzioni generalizzando dai dati di apprendimento. Quando ci muoviamo sui metodi generativi, e quindi con lo scopo non più di proporre un suggerimento scelto tra diverse opzioni, ma di generare un testo, una immagine un pezzo di codice, etc. a valle di una richiesta (*prompt*) cosa deve essere la spiegazione e quali metodi possono produrla è ancora tema di ricerca.

Inoltre, ulteriori criticità emergono in relazione ai grandi modelli generativi (*LLM*)– i modelli di IA a finalità generali, nei quali la maggiore complessità strutturale ed il regime proprietario ostacolano ulteriormente l'impiego di tecniche di *XAI* e richiedono lo sviluppo di nuove strategie di spiegabilità[6].

Il punto di vista sostenuto in questo intervento è che l'AI Act, con la rilevanza che pone sul requisito di trasparenza, trova nelle tecnologie XAI un supporto fondamentale, che deve essere incorporato nel ciclo di vita dello sviluppo ed utilizzo dei sistemi AI.

L'AI Act ne ha rafforzato la necessità e sicuramente stimolerà ulteriormente la ricerca e l'innovazione verso forme più comprensibili e affidabili di IA, volte a supportare una collaborazione sinergica tra decisori umani e macchina. Questo è coerente con l'attuale ricerca in XAI ed AI in generale, verso forme conversazionali ed interattive consapevoli dei processi cognitivi del decisore umano e del perimetro delle competenze dei modelli AI: "*io so di non sapere*".

Le sezioni successive sono volte ad evidenziare gli elementi del regolamento che esplicitamente si collegano agli aspetti operazionali che sottendono alla trasparenza e spiegabilità.

2. La normativa europea su trasparenza e spiegabilità

2.1 Trasparenza e contestabilità nei sistemi di IA ad alto rischio

Da una prospettiva legale, la spiegazione di un sistema di intelligenza artificiale non serve solo a capirne il funzionamento, ma anche a permettere alle persone di contestare il suggerimento proposto[7]. Questo concetto è legato al dibattito sul diritto di ottenere una spiegazione per le decisioni automatizzate, come previsto già dall'articolo 22 del GDPR.

L'introduzione del diritto alla spiegazione nell'articolo 86 dell'AI Act ha parzialmente risolto questa questione. Non si tratta di un nuovo diritto, ma piuttosto di un rimedio limitato, applicabile solo ad alcuni sistemi di IA ad alto rischio – escludendo, ad esempio, i sistemi medicali – da far valere in un secondo momento verso chi usa applicazioni di AI in contesti reali ricompresi nell'allegato III[8].

È importante distinguere tra spiegazione e giustificazione. Comprendere il funzionamento di un

[6] J. Schneider, *Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda.* Artificial Intelligence Review, 57, 11, 2024, 289. https://doi.org/10.1007/s10462-024-10916-x.

[7] M. Hildebrandt, *Qualification and Quantification in Machine Learning. From Explanation to Explication* in *Sociologica*, 16, 3, 2023, 37-49. https://doi.org/10.6092/ISSN.1971-8853/15845.

[8] A. Blatti, S. Tramacere, *The transparency and liability issues associated with AI- based medical systems*, in Enabling and Safeguading Personalized Medicine, in F. Casarosa, F. Gennari, A. Rossi, (a cura di), *Data Science, Machine Intelligence, and Law, forthcoming publication*, 2025, https://link.springer.com/book/9783031997082#back-to-top.

algoritmo non garantisce che le sue conclusioni siano legali. La vera rilevanza giuridica sta nell'intero processo di sviluppo e nelle scelte fatte dai creatori. È necessario poter verificare che:

- A monte, l'obiettivo del sistema sia legittimo.
- Durante lo sviluppo, sviluppatori e fornitori rispettino le normative: l'AI Act e il GDPR.
- A valle, la decisione finale sia spiegabile, interpretabile e conforme alla legge, evitando ad esempio discriminazioni illegittime[9].

La trasparenza è cruciale per un sistema di accountability efficace. Permette a tutti gli attori della filiera di sviluppo di essere "chiamati a rispondere" delle loro scelte, documentandole e giustificandole[10]. La sua funzione strumentale è esplicitata anche dall'articolo 13 dell'AI Act, che obbliga i fornitori a progettare sistemi ad alto rischio in modo trasparente per garantire il rispetto degli obblighi normativi previsti nella sezione 3.

2.2 Trasparenza e supervisione umana nei sistemi ad alto rischio nell'AI Act

L'articolo 13(1) dell'AI Act obbliga chi crea sistemi di intelligenza artificiale ad alto rischio (provider) a renderli così trasparenti da permettere a chi li usa (deployer) di interpretare il suggerimento.

In questo contesto, l'AI Act assolve ad una duplice funzione, i) riconoscere l'importanza di usare modelli di IA che si possano spiegare e di ii) spingere a sviluppare metodi di spiegazione più robusti e affidabili, superando i limiti delle attuali tecniche, specialmente per i modelli generativi.

L'articolo 14(4)(c) focalizza sulla qualità della spiegazione, chiedendo che l'output sia interpretabile "correttamente" dalle persone. Questo conferma la necessità di utilizzare forme di comunicazioni appropriate per l'utente ma anche affidabili. Questo va nella direzione di evitare soluzioni superficiali, come quelle basate sulla *Chain-of-Thought*, che possono sembrare plausibili ma non sono sempre veritiere o fedeli al funzionamento interno del modello. Non a caso, il *Code of Practice*, nel capitolo dedicato a *Safety and Security*[11], annovera il *self-reasoning* – ovvero la capacità di un modello di ragionare su se stesso e sulla propria implementazione – fra le potenziali fonti di rischio sistemico legato alla perdita di controllo umano (Appendici 1.3 e 1.4).

La trasparenza è fondamentale per una supervisione umana efficace. L'articolo 14 impone ai provider di *"progettare e sviluppare i sistemi ad alto rischio in modo tale da poter essere efficacemente supervisionati da persone fisiche durante il periodo in cui sono in uso"*. I supervisori devono poter: comprendere i limiti e le capacità del sistema; essere consapevoli del rischio di fidarsi ciecamente dell'output (il cosiddetto "*automation bias*"); interpretare correttamente l'output; decidere in ogni momento di non usare il sistema o di ignorarne i risultati.

2.3 L'importanza del *design ex ante* per tutelare la persona

In linea con la logica sottostante all'AI Act e in ossequio al principio di *accountability*, disporre di un sistema di IA, così come di una complessiva

[9] Corte di Giustizia dell'Unione Europea (CGUE), C-817/19, *Lingue des droit humans.*

[10] G. Comandé, *Intelligenza artificiale e responsabilità tra liability e accountability. Il carattere trasformativo dell'IA e il problema della responsabilità*, in A. Nuzzo, G. Olivieri (a cura di) *Analisi Giuridica dell'Economia. Algoritmi. Se li conosci, li regoli...*, 2019.

[11] EU Commission, *Code of Practice for General-Purpose AI model, Safety and Security Chapter*, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

pipeline di sviluppo, trasparente e spiegabile necessita di una collaborazione sinergica tra i vari attori coinvolti.

Queste misure tecniche devono essere implementate dal provider prima che il sistema arrivi sul mercato. In questo modo, chi usa il sistema non viene lasciato solo a gestire i rischi. Si crea un obbligo di collaborazione per proteggere al meglio la persona. In più, si evita che il supervisore umano diventi l'unico *"eroe"* o l'unico responsabile in caso di problemi[12]..

Questo rischio si vede anche nell'articolo 86(1), che dà alla persona il diritto di chiedere una spiegazione al deployer. Tuttavia, questo diritto sarà davvero efficace solo se, a monte, il provider avrà già rispettato tutti i requisiti di trasparenza, spiegabilità e supervisione umana. Se il provider non documenta e giustifica le sue scelte in modo chiaro, per il deployer sarà molto difficile adempiere ai suoi obblighi, incluso quello previsto dall'articolo 86.

3. Il ciclo di vita della AI Responsabile e le sfide aperte

Quanto illustrato nelle due sezioni precedenti è coerente con un approccio alla progettazione di sistemi di AI che sia responsabile in tutti i suoi passi. Questa visione non è nuova: sin dagli anni '90, quando si coniò il termine "Knowledge Discovery Process" c'era la consapevolezza che l'estrazione di conoscenza da dati, sia a scopi descrittivi che predittivi, ed oggi anche generativi, è un complesso processo interattivo ed iterativo composto da diverse fasi che possono coinvolgere diversi attori. La messa in esercizio di sistemi AI necessita pertanto di poter fare riferimento a standard di realizzazione, utilizzo e monitoraggio. Il primo esempio è la metodologia CRISP-DM[13] che definisce in dettaglio la documentazione di ogni scelta tecnica al fine di esplicitare le misure di qualità e sicurezza adottate ad ogni passo. Negli anni, le nuove potenzialità della tecnologia hanno reso obsolete alcune parti di quel processo ed hanno richiesto revisioni continue. Il tema del "*disegno responsabile*" è oggetto di una crescente attenzione, alla ricerca di metodologie e standard di disegno che catturino i nuovi scenari dell'AI sia predittiva/discriminativa che generativa.

Ad esempio, nel libro "Responsible AI" di Virginia Dignum[14], il ciclo di vita di un sistema AI si struttura come un processo iterativo che collega sviluppo, uso, analisi e re-design. Ogni passaggio è affiancato da strumenti di controllo e mitigazione dei rischi: il passaggio dallo sviluppo all'uso prevede procedure di "*incident response"* per intercettare e gestire tempestivamente malfunzionamenti o effetti inattesi; tra uso e analisi, l'*impact assessment* permette di valutare conseguenze sociali, etiche ed economiche; dall'analisi si passa al re-design tramite una *design review* che traduce l'evidenza empirica in miglioramenti progettuali; infine, il ritorno allo sviluppo integra il *harm modeling*, volto ad anticipare e limitare rischi futuri. In questo modo, il ciclo rafforza l'affidabilità e la capacità adattiva del sistema.

Quindi, nel caso di IA predittiva/discriminativa, su cui si baseranno la maggior parte delle applicazioni ad alto rischio, le tecniche di XAI e le metodologie di sviluppo offrono una base di partenza solida, anche se rimangono ulteriori sfide

[12] C. CASONATO, *Unlocking the synergy: Artificial intelligence and (old and new) human rights*, in *BioLaw Journal*, 3, 2023, 233-240.

[13] C. SHEARER, *The CRISP-DM Model: The New Blueprint for Data Mining* in *Journal of Data Warehousing*, 5, 2020, 13-22.

[14] V. DIGNUM, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, 2019 (1st. ed.).

di ricerca. Sarà fondamentale l'avanzamento nell'uso delle tecniche XAI per realizzare la collaborazione sinergica fra persona e macchina, incorporando una maggiore sensibilità verso i processi cognitivi che sottendono le decisioni umane. Questo comporterà anche estendere le tecniche di XAI ad esempio verso modelli capaci di astenersi dal decidere in caso di incertezza, di fornire spiegazioni causali, di sostenere un dialogo interattivo con adeguate metafore visuali o conversazionali[15]. La storia parallela delle tecniche di XAI e dell'AI Act, è una storia di reciproca influenza tra lo sviluppo dell'una e la costruzione dell'altra, e quello che possiamo aspettarci è che questa reciproca influenza aumenti nei prossimi anni con la piena entrata in vigore dell'Ai Act e dello sforzo di operazionalizzazione dei principi fondativi, *trasparenza in primis*.

Uno scenario diverso si presenta invece per l'IA generativa[9]. L'impiego di sistemi di IA generativa in applicazioni ad alto rischio rappresenta una sfida largamente aperta. La mole di dati usati per l'addestramento, l'architettura estremamente complessa e l'emergere di comportamenti non prevedibili rendono arduo anticipare le dinamiche e, di conseguenza, assicurare un comportamento affidabile[16]. In questo contesto, se da una parte sembra necessario disporre di tecniche di XAI in grado di supportare la verifica di affidabilità, la supervisione umana, l'accountability, dall'altra parte c'è la consapevolezza che le tecniche sviluppate fino ad ora non si adattano "*as is*" ed occorre sviluppare forme di XAI completamente nuove ed è in corso una vibrante attività di ricerca scientifica e tecnologica a riguardo. L'AI Act rappresenta pertanto uno stimolo fondamentale alla ricerca di *IA generativa responsabile*, *sicura e al servizio delle persone*.

[15] C. Punzi, R. Pellegrini, M. Setzu, F. Giannotti, D. Pedreschi, *Ai, meet human: Learning paradigms for hybrid decision making systems*, arXiv, 2024: https://arxiv.org/abs/2402.06287.

[16] Y. Huang et al., *Trustllm: Trustworthiness in large language models*, arXiv, 2024: https://arxiv.org/abs/2401.05561.