Rischi etico-legali dell'Intelligenza Artificiale

di Anna Monreale

Abstract: Ethical and legal risks of Artificial Intelligence – Nowadays, a wide variety of personal data, describing directly or indirectly individuals' activities, are collected and available, due to the spread of different apps, sensors and mobile devices. The availability of such data opens unprecedented opportunities of developing AI systems exploiting those data to provide a wide range of benefits in different domains. Unfortunately, in a critical domain like healthcare and justice the development and the application of AI systems can rise many ethical and legal concerns. This article discusses the ethical and legal implication in terms of privacy and transparency, also providing an overview on the different approaches to ethics across the world.

Keywords: Artificial intelligence; Ethics; Privacy; Transparency; Interpretability.

1. Introduzione

L'opportunità di poter definire e sviluppare modelli di Intelligenza Artificiale (IA), capaci di apprendere dai dati regole che potrebbero supportare il processo decisionale umano, ha dato spazio a benefici e potenzialità senza precedenti in domini applicativi diversi. Questa grande opportunità è garantita dalla disponibilità di dati che descrivono nel dettaglio miriadi di attività e fenomeni del mondo reale. Esempi di domini applicativi che potrebbero usufruire dell'uso di sistemi di IA sono l'assistenza medico-sanitaria, i mercati bancari e finanziari, il commercio elettronico, i social media, la giustizia, ecc.

L'efficacia di sistemi di IA basati sul Machine Learning è stata dimostrata più e più volte. Gli algoritmi di machine learning costruiscono modelli predittivi in grado di mappare le caratteristiche dell'utente a una classe (o decisione) grazie a una fase di apprendimento. Questo processo di apprendimento è reso possibile dalle tracce digitali, che le persone si lasciano alle spalle mentre svolgono le proprie attività giornaliere (ad es. movimenti, acquisti, commenti nei social network, ecc.). Questa enorme quantità di dati può contenere pregiudizi umani ma anche informazioni molto sensibili che descrivono aspetti privati della sfera personale. Pertanto, i modelli decisionali appresi su di essi possono ereditare tali pregiudizi, portando probabilmente a decisioni sbagliate e ingiuste, oppure possono violare la privacy delle persone le cui abitudini sono descritte dai dati.

I modelli decisionali sfruttano spesso algoritmi molto sofisticati che permettono di ottenere delle ottime prestazioni in termini di accuratezza dei

risultati; purtroppo le ottime prestazioni spesso sono ottenute con modelli molto complessi, come *Neural Network* e *Deep Neural Network*, che sono di difficile interpretazione anche per gli esperti del settore. In altre parole, è difficile comprendere il loro comportamento interno in modo da giustificare una possibile decisione. Questo porta all'uso di tali sistemi come una *scatola chiusa* di cui bisogna fidarsi ciecamente senza la possibilità di scrutarci dentro per una maggiore comprensione. Chiaramente, in campi in cui la comprensione del meccanismo decisionale diventa di fondamentale importanza (es. campo medico, campo giuridico) l'applicabilità di tali sistemi diventa critica. Questo corrisponde a chiedere ad esperti, come per esempio medici e giudici, di prendere decisioni, che potrebbero avere un forte impatto sulla vita di una persona, fidandosi di un sistema che non si conosce e comprende. Una situazione del genere è veramente lontana dall'idea di un'Intelligenza Artificiale capace di collaborare con l'essere umano per migliorare le sue abilità, mentre è più vicina a un'idea d'Intelligenza Artificiale che tende a sostituire l'uomo. Quindi, sebbene questi modelli decisionali siano molto efficaci ed efficienti, non è possibile ignorare alcune delle implicazioni etico-legali che possono derivare dal loro uso come: violazione del diritto alla privacy, violazione del diritto alla spiegazione e mancanza di trasparenza.

L'importanza di identificare delle linee guida e dei requisiti di un'Intelligenza Artificiale responsabile e affidabile è stata riconosciuta dalle maggiori potenze mondiali. Questo articolo presenta una breve analisi del sistema dei valori delle tre potenze più influenti del mondo (Europa, USA e Cina) e del loro approccio rispetto agli aspetti etico-legali relativi alle tecnologie di Intelligenza Artificiale. L'analisi evidenzia come questi paesi siano guidati da ideologie completamente diverse che li portano ad avere un approccio diverso all'innovazione basata sull'intelligenza artificiale. Infine, l'articolo presenta un'analisi più focalizzata sui requisiti di privacy e trasparenza che rappresentano i rischi che maggiormente bloccano l'adozione e la diffusione dell'Intelligenza Artificiale.

2. L'approccio all'etica dell'IA nel mondo

Le linee guida etico-legali per l'uso responsabile dell'IA variano in tutto il mondo. Negli ultimi anni, sono stati prodotti diversi report che descrivono e discutono i requisiti di un'implementazione affidabile delle tecnologie di intelligenza artificiale. Dalla letteratura emergono i seguenti principi etici: *trasparenza*, *giustizia*, *non maleficenza*, *responsabilità* e *privacy*. Comunque, ogni paese interpreta questi principi, li traduce e li adegua al proprio sistema legale.

La principale dimensione rispetto a cui i diversi approcci differiscono riguarda il bilanciamento tra *regolamentazione* e *innovazione*. Per esempio, gli USA tendono a favorire l'innovazione, mentre l'Unione Europea segue un approccio prevalentemente guidato dalla regolamentazione. Di conseguenza, i problemi etico-legali che possono scaturire dallo sviluppo di tecnologie di intelligenza artificiale possono variare da paese a paese. Di seguito presentiamo una breve analisi del sistema dei valori delle tre potenze più influenti del mondo, che hanno

emanato delle politiche per un'intelligenza artificiale responsabile. In particolare, evidenziamo come questi paesi sono guidati da ideologie completamente diverse che li portano ad avere un approccio diverso all'innovazione basata sull'intelligenza artificiale.

USA. L'approccio americano agli aspetti etici dell'IA è influenzato dai valori libertari che implicano una regolamentazione minima della tecnologia da parte del governo; in particolare, gli USA promuovono un "modello Silicon Valley"[1] che si basa sul principio "*move fast, break things first, apologize later*". In altre parole, il sistema è basato sulla concezione umana di *Homo Economicus* e dell'*individualismo.* Nel 2020, l'amministrazione di Trump ha pubblicato una bozza della guida per la regolamentazione dell'applicazione dell'IA,[2] che scoraggia ogni azione che ostacoli l'innovazione e la crescita, coerentemente con l'approccio americano di dare priorità all'innovazione piuttosto che alla legge.

Cina. L'approccio cinese è invece influenzato dai valori confuciani e dall'ideologia del socialismo cinese. Il sistema è caratterizzato da un'attenzione particolare all'armonia sociale che implica alcuni elementi di controllo morale e sorveglianza da parte del governo. L'approccio si basa su una concezione dell'essere umano di tipo *collettivista* e *utilitarista.* La Cina nel 2017 ha definito un piano strategico di sviluppo rispetto all'IA "*The Next Generation Artificial Intelligence Development Plan*", che ha lo scopo di definire gli obiettivi strategici, i task principali e le misure per il supporto dello sviluppo di dell'IA entro il 2030.

Unione Europea. L'approccio europeo è invece basato sul rispetto dei diritti fondamentali, della democrazia e della legge. In particolare, quattro principi etici sono considerati molto rilevanti per l'IA: rispetto per l'autonomia umana, prevenzione da rischi, equità e spiegazione. Il sistema europeo si basa sulla concezione kantiana della persona come autonoma (libertà, autonomia e dignità). L'Unione Europea comunque è probabilmente il leader mondiale nella regolamentazione dei principi etici dell'IA e nell'influenza della discussione internazionale su questo argomento.[3]

La propensione dell'Unione Europea a codificare in modo rigido i suoi principi etici sull'Intelligenza Artificiale ha sollevato alcune preoccupazioni sul fatto che quest'approccio potrebbe costituire un ostacolo all'innovazione. Tuttavia, la Commissione Europea vede la codifica dei principi etici nell'Intelligenza Artificiale come un vantaggio competitivo che promuoverà la fiducia dei consumatori nei prodotti dell'Unione Europea. Infatti, la Commissione Europea promuove lo sviluppo di tecnologia affidabile e ne vuole armonizzare l'adozione in

[1] A. Armitage, A. Cordova, and R. Siegel, *Design-thinking: The answer to theimpasse between innovation and regulation*, UC Hastings Research Paper. (250) (2017)

[2] White House's Officeof Scienceand Technology Policy. *Guidance for regulation of artificial intelligence applications.* www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf (2019).

[3] *Dichiarazione IBM sulle linee guida etiche dell'UE per l'IA affidabile.* www.ibm.com/blog/politica/AI-etica-eu

tutta l'Unione. Un esempio recente di regolamentazione che ha un impatto anche sulle applicazioni di IA e il regolamento sulla protezione dei dati personali.[4]

La Commissione Europea ha anche nominato un gruppo di esperti chiamato "*High-Level Expert Group on AI*" (AI HLEG) per la definizione di una strategia sull'IA. Nel 2019, tale gruppo ha pubblicato delle line guida che indicano sette requisiti per lo sviluppo di strumenti di IA responsabile[5].

L'essenza dell'approccio europeo è la promozione della ricerca e l'innovazione che agisce in modo responsabile, cioè mette in atto tutte le protezioni necessarie a garantire la protezione dei diritti e la libertà dei cittadini europei.

Tra i sette principi identificati dall'AI HLEG troviamo la *privacy* e *data governance*, e la *trasparenza* che include l'*interpretabilità* dell'intelligenza artificiale. Tali argomenti verranno trattati esplicitamente di seguito anche dal punto di vista tecno-scientifico.

3. *Privacy* e *Data Governance*

Tale requisito è in linea con gli Articoli 7 e 8 della Carta dei Diritti Fondamentali dell'Unione Europea sul "*Rispetto della vita privata e familiare*" e sulla "*Protezione dei dati personali*", i quali riflettono il principio della prevenzione dei rischi applicato alla privacy. La protezione dei dati personali è inoltre regolata dal GDPR, insieme ad altre direttive diffuse in tutta l'Unione Europea. Le linee guida prescrivono un'attenzione particolare per i dati sensibili che includono orientamento religioso, sessuale e politico, età e genere, informazioni che potrebbero essere dedotte dalle tracce digitali degli utenti e usati per discriminarli. Per essere conforme a questo requisito, è necessario considerare due aspetti chiave:

- *Privacy* e protezione dei dati personali. Garantire la protezione dei dati sensibili come quelli sanitari durante l'intero ciclo di vita della tecnologia basata sull'IA è fondamentale per agevolare l'adozione e la diffusione di questi sistemi. Chiaramente questo include anche impostare ed eseguire un meccanismo di valutazione sistematica che riesca a quantificare l'impatto di ogni sistema sulla protezione dei dati.
- *Data governance*: accesso, qualità e integrità dei dati. La data governance è il processo di gestione dei dati utilizzati da un'organizzazione. Tutto questo include l'implementazione di protocolli per l'accesso ai dati, procedure per garantire la qualità dei dati, cioè che permettono di ottenere dati privi di distorsioni, di errori di qualsiasi natura, e meccanismi per la valutazione dell'integrità dei dati.

La protezione dei dati personali diventa di fondamentale importanza quando gli algoritmi di machine learning sono addestrati su dati sensibili come quelli sanitari. In questi casi è necessario applicare il principio della *privacy by design*

[4] EU General Data Protection Regulation: gdpr-info.eu
[5] High-Level Expert Group on AI. *The Ethics Guidelines for Trustworthy Artificial Intelligence (AI)*. ec.europa.eu/futurium/en/ai-alliance-consultation (04, 2019)

introdotto da Ann Cavoukian nel 1990[6] è successivamente elaborato da Monreale et al.[7] L'idea principale è identificare e considerare i requisiti di privacy nella progettazione del processo di estrazione di conoscenza, che permette l'apprendimento del modello di machine learning, allo scopo di garantire la protezione di dati personali e ottenere modelli accurati. I rischi di privacy possono derivare sia dall'accesso a dati sensibili durante la fase di apprendimento che dall'utilizzo e quindi l'accesso al modello di machine learning previsionale.

L'apprendimento dei modelli di machine learning viene eseguito su dati de-identificati, cioè dati che non identificano direttamente una persona in modo univoco. Purtroppo, la de-identificazione in molti contesti non è sufficiente a proteggere la privacy, poiché dati completamente pubblici possono essere correlati con i dati de-identificati, portando così alla re-identificazione indiretta delle persone[8]. L'obiettivo delle tecniche per la tutela della privacy, come il *k*-anonimato[9] e *l*-diversity[10], è quindi quello di trovare contromisure a questo tipo di attacco, in modo tale che la capacità dell'avversario di collegare i dati resi anonimi con altre informazioni (quasi-identificatori) è limitata.

Alcune violazioni della privacy possono derivare invece da attacchi che operano direttamente sul modello di machine learning e cercano di inferire la presenza o meno di alcuni individui nei dati che sono stati usati per l'apprendimento del modello. Questo tipo di attacco è chiamato "*membership inference attack*"[11]. Questo attacco assume che l'avversario abbia solo la possibilità di interrogare il modello di machine learning e di ottenere la decisione su un insieme di istanze da valutare. Partendo da queste informazioni è possibile apprende un modello di machine learning capace di distinguere tra il comportamento del modello sotto attacco nei confronti di individui usati per l'apprendimento e individui non usati per apprendere il predittore. La maggior parte delle strategie per garantire la protezione contro violazioni della privacy di questo tipo si basano su metodi che sfruttano la "*differential privacy*", un modello di privacy basato su un concetto di perturbazione di dati o funzioni[12].

---

[6] A. Cavoukian, *Privacy design principles for an integrated justice system.* Working paper. (2000). www.ipc.on.ca/index.asp?layid=86%26;fid1=318.

[7] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi, *Privacy-by-design in big data analytics and social mining*, EPJ Data Science. 3(1), 10 (2014).

[8] H. Jones, *Geoff hinton dismissed the need for explainable AI: 8 experts explain why he's wrong*, Forbes, Dec. 20 (2018).

[9] L.Sweeney, *k-anonymity: A model for protecting privacy*, Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 10(5), 557–570 (2002).

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. *l- diversity: Privacy beyond k-anonymity.* In eds. L. Liu, A. Reuter, K. Whang, and J. Zhang, Int. Conference on Data Engineering, ICDE, p. 24 (2006).

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. *Membership inference attacks against machine learning models.* In IEEE Symposium on Security and Privacy, pp. 3–18 (2017).

[12] C. Dwork. *Differential privacy.* In eds. M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Proceedings, Part II, vol. 4052, Lecture Notes in Computer Science, pp. 1–12, Springer (2006).

4. Trasparenza e Interpretabilità

I requisiti di trasparenza e interpretabilità sono esplicitamente richiesti dalle linee guida dell'Unione Europea sull'IA. La trasparenza dovrebbe essere considerata ad ogni fase del ciclo di vita della tecnologia di IA. Trasparenza significa avere la possibilità di ottenere una visione completa di tutto il sistema in ogni momento. Per essere conformi a questo requisito, tutti i passi necessari per l'implementazione del sistema di intelligenza artificiale dovrebbero essere propriamente documentati. Questo include documentare la fase di raccolta dati, le strategie algoritmiche e le architetture usate, documentare come i dati sono stati usati per l'apprendimento e per la fase di validazione e test del sistema. Inoltre, dovrebbero essere forniti tutti gli strumenti per poter comprendere il sistema decisionale e il ragionamento usato per prendere le decisioni. Quest'ultimo aspetto è di particolare importanza per quei sistemi di intelligenza artificiale le cui decisioni hanno un impatto sulla vita delle persone. Rendere comprensibile il ragionamento dei sistemi decisionali basato sull'IA è uno aspetto ritenuto fondamentale per la definizione e realizzazione di un'intelligenza artificiale responsabile e affidabile, ma è anche un requisito richiesto dal GDPR (Art. 22). Purtroppo, la maggior parte dei sistemi di IA sfruttano modelli di machine learning molto complessi che non sono trasparenti e comprensibili neanche ad esperti del settore poiché la ragione della decisione non può essere interpretata semplicemente guardando i parametri interni del sistema.

Esistono due modi per poter ottenere il livello di trasparenza richiesto per legge: *i)* usare modelli di IA interpretabili oppure *ii)* usare delle tecniche che permettono di fornire *spiegazioni* comprensibili sul comportamento di un sistema decisionale che sfrutta modelli non interpretabili.

Modelli interpretabili. Dire che un modello è interpretabile significa che è capace di trasmettere una spiegazione comprensibile agli esseri umani. Quindi, l'uomo deve essere capace di capire e interpretare come un determinato dato in input al sistema è matematicamente mappato a una determinata decisione. Esistono pochissimi modelli che sono considerati direttamente interpretabili e sono: gli alberi decisionali, i modelli lineari e le regole di classificazione[13]. Un albero decisionale sfrutta un grafo strutturato come un albero e composto da nodi interni che rappresentano test sulle variabili (ad esempio, controllano se una variabile ha un valore inferiore, uguale o maggiore di una soglia) e nodi foglia che rappresentano una decisione. Ogni ramo rappresenta una possibile decisione. I percorsi dalla radice alle foglie rappresentano le *regole di classificazione*. Le regole più comuni sono le regole di tipo *if-then*, in cui la clausola "if" è una combinazione di condizioni sulle variabili di input. Se la clausola è verificata, la parte "allora" rivela l'azione da intraprendere o la decisione. I modelli lineari invece consentono di identificare l'importanza delle variabili, che descrivono un'istanza, ai fini della sua classificazione.

Sebbene in molti casi questi modelli raggiungono buoni risultati in termini di accuratezza nel processo predittivo, essi non possono essere usati in caso sia

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, *A survey of methods for explaining black box models*, ACM Comput. Surv. 51(5), 93:1–93:42 (2019).

necessario costruire modelli predittivi basati su dati non di tipo tabellare. Infatti, nel caso in cui i dati da cui effettuare l'apprendimento siano immagini, serie temporali e testi i modelli interpretabili richiedono una fase di definizione e estrazione di variabili da derivare dal dato originale, poiché altrimenti tali modelli non sarebbero direttamente applicabili.

Modelli interpretabili per la spiegazione dell'IA. Spesso i modelli decisionali, che permettono di raggiungere livelli di accuratezza rilevanti e accettabili, sono molto complessi, quindi in questi casi è necessario applicare delle tecniche che permetto di fornire spiegazioni sul loro funzionamento interno comprensibili all'essere umano. Tipicamente, queste tecniche analizzano il comportamento del modello non interpretabile e cercano di imitarlo usando modelli interpretabili.

Esistono due macro-famiglie di tecniche di spiegazione: tecniche *globali* e tecniche *locali* di spiegazione. Le prime cercano di derivare un modello interpretabile che descrive l'intero modello di IA originale nella sua complessità. Purtroppo, queste tecniche tendono a generare dei modelli come alberi decisionali caratterizzati da una grandissima quantità di nodi, quindi perdendo la loro immediata interpretabilità. Le tecniche di spiegazione locali invece si pongono l'obiettivo di spiegare la ragione di una singola decisione. Quindi cercano di imparare un modello interpretabile che approssima il comportamento del modello complesso nel vicinato dell'istanza su cui è necessario prendere una decisione. Questi metodi risultano molto efficaci poiché approssimare il comportamento locale del modello complesso è molto più semplice rispetto ad imitare l'intero suo comportamento. La comunità scientifica ha sviluppato differenti metodi basati su questo approccio generale capaci di spiegare decisioni su diverse forme di dati: immagini, testi, e tabelle.[14,15,16]

5. Conclusioni

I sistemi di Intelligenza Artificiale hanno il potere di migliorare i processi decisionali di molti settori, come quello medico-sanitario, economico-finanziario, giuridico, grazie a modelli di machine learning e deep learning sempre più efficaci. In questo articolo, abbiamo discusso possibili implicazioni etico-legali dello sviluppo e della diffusione di tali tecnologie. Abbiamo fornito una panoramica dei diversi approcci all'etica dell'Intelligenza Artificiale da parte delle maggiori potenze mondiali e infine abbiamo analizzato anche dal punto di vista tecno-scientifico i requisiti di privacy e trasparenza nei modelli di Intelligenza Artificiale, evidenziando che l'approccio all'etica in tutto il mondo è profondamente eterogeneo. Nell'era della globalizzazione, questo potrebbe essere un ostacolo

[14] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. *Factual and counterfactual explanations for black box decision making*. IEEE Intelligent Systems, 34(6):14–23, 2019.
[15] R. Guidotti, A. Monreale, S. Matwin, and D. Pedreschi. *Black box explanation by learning image exemplars in the latent feature space*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 189–205. Springer, 2019.
[16] S. M. Lundberg and S.-I. Lee. *A unified approach to interpreting model predictions*. In Advances in neural information processing systems, pages 4765–4774, 2017.

Saggi – *DPCE online*, 2020/3
ISSN: 2037-6677

all'armonizzazione normativa di queste tecnologie che a sua volta può portare a un mercato globale frammentato. Inoltre, ciò potrebbe minare la fiducia nei sistemi di intelligenza artificiale e rallentarne l'adozione in scenari reali.

Anna Monreale
Dip.to Informatica
Università di Pisa
anna.monreale@unipi.it