AI spiegabile: le basi

INFORMATIVA SULLA POLITICA

THE ROYAL SOCIETY

AI spiegabile: le basi

Informazioni sulla politica

Rilasciato: novembre 2019 DES6051

ISBN: 978-1-78252-433-5

Questo rapporto può essere visualizzato online all'indirizzo:

royalsociety.org/ai-interpretability

Immagine di copertina © maxkabakov.

Contenuti

Riepilogo

Gli ultimi anni hanno visto progressi significativi nelle capacità delle tecnologie di intelligenza artificiale (AI). Molte persone ora interagiscono quotidianamente con i sistemi abilitati all'intelligenza artificiale: nei sistemi di riconoscimento delle immagini, come quelli utilizzati per taggare le foto sui social media; nei sistemi di riconoscimento vocale, come quelli utilizzati dagli assistenti personali virtuali; e nei sistemi di raccomandazione, come quelli utilizzati dai rivenditori online.

Man mano che le tecnologie di intelligenza artificiale diventano integrate nei processi decisionali, si è discusso nelle comunità di ricerca e politiche sulla misura in cui gli individui

sviluppando l'IA, o soggetti a una decisione abilitata all'IA, sono in grado di capire come funziona il sistema decisionale risultante.

Alcuni degli strumenti di intelligenza artificiale odierni sono in grado di produrre risultati altamente accurati, ma sono anche molto complessi. Questi cosiddetti modelli "scatola nera" possono essere troppo complicati per essere compresi appieno anche da utenti esperti. Poiché questi sistemi vengono implementati su larga scala, ricercatori e responsabili politici si chiedono se l'accuratezza in un compito specifico superi altri criteri importanti nei sistemi decisionali. I dibattiti politici in tutto il mondo vedono sempre più la richiesta di una qualche forma di spiegabilità dell'IA, come parte degli sforzi per incorporare principi etici nella progettazione e nell'implementazione di sistemi abilitati all'IA. Questo briefing si propone quindi di riassumere alcuni dei

problemi e considerazioni durante lo sviluppo di metodi di intelligenza artificiale spiegabili.

Ci sono molte ragioni per cui una qualche forma di interpretabilità nei sistemi di IA potrebbe essere desiderabile o necessaria. Questi includono: dare agli utenti la certezza che un sistema di intelligenza artificiale funziona bene; protezione contro i pregiudizi; aderire a standard normativi o requisiti politici; aiutare gli sviluppatori a capire perché un sistema funziona in un certo modo, valutarne le vulnerabilità o verificarne gli output; o soddisfare le aspettative della società su come agli individui viene offerta la possibilità di agire in un processo decisionale.

Diversi metodi di intelligenza artificiale sono influenzati dalle preoccupazioni sulla spiegabilità in modi diversi. Proprio come esiste una gamma di metodi di intelligenza artificiale, esiste anche una gamma di approcci alla spiegabilità. Questi approcci svolgono funzioni diverse, che possono essere più o meno utili, a seconda dell'applicazione in uso. Per alcune applicazioni può essere possibile utilizzare un sistema interpretabile in base alla progettazione, senza sacrificare altre qualità, come la precisione.

Ci sono anche insidie associate a questi diversi metodi e coloro che utilizzano i sistemi di intelligenza artificiale devono considerare se le spiegazioni che forniscono sono affidabili, se c'è il rischio che le spiegazioni possano trarre in inganno i loro utenti o se potrebbero contribuire al gioco del sistema o alle opportunità per sfruttarne le vulnerabilità.

Contesti diversi danno origine a diverse esigenze di spiegabilità e la progettazione del sistema spesso deve bilanciare esigenze concorrenti, ad esempio per ottimizzare l'accuratezza di un sistema o garantire la privacy degli utenti. Esistono esempi di sistemi di intelligenza artificiale che possono essere implementati senza dare adito a preoccupazioni sulla spiegabilità, generalmente in aree in cui non ci sono conseguenze significative da risultati inaccettabili o il sistema è ben convalidato. In altri casi, una spiegazione su come funziona un sistema di IA è necessaria ma potrebbe non essere sufficiente per dare fiducia agli utenti o supportare meccanismi efficaci di responsabilità.

In molti sistemi decisionali umani, nel tempo si sono sviluppati processi complessi per fornire salvaguardie, funzioni di audit o altre forme di responsabilità. La trasparenza e la spiegabilità dei metodi di IA possono quindi essere solo il primo passo nella creazione di sistemi affidabili e, in alcune circostanze, la creazione di sistemi spiegabili può richiedere sia questi approcci tecnici che altre misure, come la garanzia di determinate proprietà. Coloro che progettano e implementano l'IA devono quindi considerare come il suo utilizzo si adatti al più ampio contesto sociale

contesto tecnico del suo dispiegamento.

L'IA e la "scatola nera"

Il problema di spiegabilità dell'IA

AI è un termine generico. Si riferisce ad una suite di tecnologie in cui i sistemi informatici sono programmati per esibire comportamenti complessi - comportamenti che in genere richiederebbero intelligenza negli esseri umani o negli animali - quando agiscono in ambienti difficili.

Gli ultimi anni hanno visto progressi significativi nelle capacità delle tecnologie di intelligenza artificiale, come risultato degli sviluppi tecnici nel campo, in particolare nell'apprendimento automatico1 ; maggiore disponibilità di dati; e maggiore potenza di calcolo. Come risultato di questi progressi, i sistemi che solo pochi anni fa hanno lottato per ottenere risultati accurati possono ora superare gli esseri umani in alcuni compiti specifici2.

Molte persone ora interagiscono quotidianamente con i sistemi abilitati all'intelligenza artificiale: nei sistemi di riconoscimento delle immagini, come quelli utilizzati per taggare le foto sui social media; nei sistemi di riconoscimento vocale, come quelli utilizzati dagli assistenti personali virtuali; e nei sistemi di raccomandazione, come quelli utilizzati dai rivenditori online.

Ulteriori applicazioni dell'apprendimento automatico sono già in fase di sviluppo in una vasta gamma di campi. Nel settore sanitario, l'apprendimento automatico sta creando sistemi che possono aiutare i medici a fornire diagnosi più accurate o efficaci per determinate condizioni. Nei trasporti sostiene lo sviluppo di veicoli autonomi e contribuisce a rendere più efficienti le reti di trasporto esistenti. Per i servizi pubblici ha il potenziale per indirizzare il sostegno in modo più efficace a chi ne ha bisogno o per adattare i servizi agli utenti3 .

Allo stesso tempo, le tecnologie di intelligenza artificiale vengono impiegate in aree politiche altamente sensibili, ad esempio il riconoscimento facciale nelle attività di polizia o nella previsione della recidiva nel sistema di giustizia penale, e aree in cui sono all'opera complesse forze sociali e politiche. Le tecnologie di intelligenza artificiale vengono quindi integrate in una serie di processi decisionali.

Da tempo c'è un dibattito crescente nelle comunità di ricerca e politiche
sulla misura in cui gli individui
sviluppando l'IA, o soggetti a una decisione abilitata all'IA, sono in grado di capire come funziona l'IA e perché è stata presa una decisione particolare4 .
Queste discussioni sono state portate in netto rilievo dopo l'adozione del regolamento generale europeo sulla protezione dei dati, che ha suscitato il dibattito sul fatto che le persone avessero o meno un "diritto a una spiegazione".

Questo briefing si propone di riassumere i problemi e le domande che sorgono quando gli sviluppatori e i responsabili politici hanno deciso di creare sistemi di IA spiegabili.

1. L'apprendimento automatico è la tecnologia che consente ai sistemi informatici di apprendere direttamente dai dati.

2. Va notato, tuttavia, che questi compiti di riferimento tendono ad essere di natura vincolata. Nel 2015, ad esempio, i ricercatori hanno creato un sistema che ha superato le capacità umane in una gamma ristretta di compiti relativi alla vista, incentrati sul riconoscimento delle singole cifre scritte a mano. Vedi: Markoff J. (2015) Un progresso nell'apprendimento dell'intelligenza artificiale rivaleggia con le capacità umane. New York Times. 10 dicembre 2015.

3. Royal Society (2017) Apprendimento automatico: il potere e la promessa dei computer che imparano con l'esempio, disponibile su www.royalsociety.org/machine-learning

4. Pasquale, F. (2015) The Black Box Society: The Secret Algorithms that Control Money and Information, Harvard University Press, Cambridge, Massachusetts

SCATOLA 1

Intelligenza artificiale, machine learning e statistica: connessioni tra questi campi

L'etichetta "intelligenza artificiale" descrive una suite di tecnologie che cercano di svolgere compiti solitamente associati all'intelligenza umana o animale. John McCarthy, che coniò il termine nel 1955, lo definì come “la scienza e l'ingegneria per costruire macchine intelligenti”; da allora sono state proposte molte definizioni diverse.

L'apprendimento automatico è una branca dell'IA che consente ai sistemi informatici di eseguire compiti specifici in modo intelligente. Gli approcci tradizionali alla programmazione si basano su regole codificate, che stabiliscono come risolvere un problema, passo dopo passo.
Al contrario, ai sistemi di apprendimento automatico viene assegnata un'attività e viene fornita una grande quantità di dati da utilizzare come esempi (e non) di come questa attività può essere raggiunta o da cui rilevare schemi. Il sistema impara quindi come ottenere al meglio l'output desiderato.
Esistono tre rami chiave dell'apprendimento automatico:

- Nell'apprendimento automatico supervisionato, un sistema viene addestrato con dati che sono stati etichettati.
  Le etichette classificano ogni punto dati in uno o più gruppi, ad esempio "mele" o "arance". Il sistema apprende come sono strutturati questi dati, noti come dati di addestramento, e li utilizza per prevedere le categorie di dati nuovi o di "test".

- L'apprendimento senza supervisione è l'apprendimento senza etichette. Ha lo scopo di rilevare le caratteristiche che rendono i punti dati più o meno simili tra loro, ad esempio creando cluster e assegnando dati a questi cluster.

- L'apprendimento per rinforzo si concentra sull'apprendimento dall'esperienza. In un tipico ambiente di apprendimento per rinforzo, un agente interagisce con il suo ambiente e gli viene assegnata una funzione di ricompensa che cerca di ottimizzare, ad esempio il sistema potrebbe essere ricompensato per aver vinto una partita. L'obiettivo dell'agente è apprendere le conseguenze delle sue decisioni, ad esempio quali mosse sono state importanti per vincere una partita, e utilizzare questo apprendimento per trovare strategie che massimizzino le sue ricompense.

Pur non avvicinandosi all'intelligenza generale a livello umano che è spesso associata al termine IA, la capacità di apprendere dai dati aumenta il numero e la complessità delle funzioni che i sistemi di apprendimento automatico possono svolgere. I rapidi progressi nell'apprendimento automatico stanno oggi supportando un'ampia gamma di applicazioni, molte delle quali le persone incontrano quotidianamente, portando a discussioni e dibattiti attuali sull'impatto dell'IA sulla società.

Molte delle idee che inquadrano i sistemi di apprendimento automatico odierni non sono nuove; le basi statistiche del campo risalgono a molti decenni fa e dagli anni '50 i ricercatori hanno creato algoritmi di apprendimento automatico con vari livelli di sofisticatezza.

L'apprendimento automatico prevede che i computer elaborano una grande quantità di dati per prevedere i risultati. Gli approcci statistici possono informare su come i sistemi di apprendimento automatico gestiscono le probabilità o l'incertezza nel processo decisionale.

Tuttavia, le statistiche includono anche aree di studio che non riguardano la creazione di algoritmi in grado di apprendere dai dati per fare previsioni o decisioni. Sebbene molti concetti fondamentali dell'apprendimento automatico abbiano le loro radici nella scienza dei dati e nella statistica, alcune delle sue capacità analitiche avanzate non si sovrappongono naturalmente a queste discipline.

Altri approcci all'IA utilizzano approcci simbolici, piuttosto che statistici. Questi approcci utilizzano la logica e l'inferenza per creare rappresentazioni di una sfida e per elaborare una soluzione.

Questo documento utilizza il termine generico "AI", pur riconoscendo che questo comprende un'ampia gamma di campi di ricerca e gran parte del recente interesse per l'IA è stato guidato dai progressi nell'apprendimento automatico.

La "scatola nera" nei dibattiti politici e di ricerca Alcuni degli strumenti di intelligenza artificiale odierni sono in grado di produrre risultati altamente accurati, ma sono anche molto complessi se non addirittura opachi, rendendo il loro funzionamento difficile da interpretare. Questi cosiddetti modelli a "scatola nera" possono essere troppo complicati per essere compresi appieno anche da utenti esperti5. Poiché questi sistemi vengono implementati su larga scala, ricercatori e responsabili politici si chiedono se l'accuratezza in un compito specifico superi altri criteri importanti nel processo decisionale6.

I dibattiti politici in tutto il mondo presentano sempre più richieste di una qualche forma di spiegabilità dell'IA, come parte degli sforzi per incorporare principi etici nella progettazione e nell'implementazione di sistemi abilitati all'IA7. Nel Regno Unito, ad esempio, tali appelli sono arrivate dal Comitato AI della Camera dei Lord, che ha affermato che "lo sviluppo di sistemi di intelligenza artificiale intelligibili è una necessità fondamentale se l'IA vuole diventare uno strumento integrale e affidabile nella nostra società"8 . Il gruppo di alto livello dell'UE sull'IA ha chiesto un ulteriore lavoro per definire i percorsi per raggiungere la spiegabilità9 ; e negli Stati Uniti, la Defense Advanced Research Projects Agency sostiene un importante programma di ricerca che cerca di creare un'IA più spiegabile10 . Poiché i metodi di intelligenza artificiale vengono applicati per affrontare le sfide in un'ampia gamma di aree politiche complesse, poiché i professionisti lavorano sempre più a fianco di strumenti decisionali abilitati all'intelligenza artificiale, ad esempio in medicina, e poiché i cittadini incontrano più frequentemente i sistemi di intelligenza artificiale in settori in cui le decisioni hanno un significato impatto, questi dibattiti diventeranno più urgenti.

La ricerca sull'IA, nel frattempo, continua a progredire a ritmo sostenuto. L'IA spiegabile è un vivace campo di ricerca, con molti metodi diversi che stanno emergendo e diversi approcci all'IA sono influenzati da queste preoccupazioni in modi diversi.

Terminologia

In queste ricerche, dibattiti pubblici e politici, viene utilizzata una serie di termini per descrivere alcune caratteristiche desiderate di un'IA. Questi includono:

- interpretabile, che implica un certo senso di capire come funziona la tecnologia;
- spiegabile, implicando che una gamma più ampia di utenti può capire perché o come è stata raggiunta una conclusione;
- trasparente, che implica un certo livello di accessibilità ai dati o all'algoritmo;
- giustificabile, implicando che vi sia una comprensione del caso a sostegno di un esito particolare; o
- contestabile, implicando che gli utenti abbiano il informazioni di cui hanno bisogno per argomentare contro una decisione o una classificazione.

5. Come osserva Rudin (2019), questo termine si riferisce anche a modelli proprietari a cui è negato l'accesso agli utenti. Rudin, C. (2019) Smetti di spiegare i modelli di apprendimento automatico della scatola nera per decisioni ad alto rischio e usa invece modelli interpretabili, Intelligenza della macchina della natura, 1, 206-215

6. Doshi-Velez F. e Kim B. (2018) Considerazioni per la valutazione e la generalizzazione nell'apprendimento automatico interpretabile. In: Escalante H. et al. (a cura di) Modelli spiegabili e interpretabili in Computer Vision e Machine Learning. La serie Springer sulle sfide nell'apprendimento automatico. Springer

7. Si veda l'allegato 1 per uno schizzo del panorama politico

8. House of Lords (2018) AI nel Regno Unito: pronta, disponibile e capace? Report of Session 2017 – 19. HL Paper 100.

9. Gruppo di alto livello dell'UE sull'IA (2019) Linee guida etiche per un'IA affidabile. Disponibile su: https://ec.europa.eu/digital-single market/en/news/ethics-guidelines-trustworthy-ai [visitato il 2 agosto 2018]

10. DARPA Explainable Artificial Intelligence (programma XAI), disponibile su: https://www.darpa.mil/program/explainable intelligenza artificiale [accesso 2 agosto 2018]

Sebbene l'uso di questi termini sia incoerente11, ognuno cerca di trasmettere il senso di un sistema che può essere spiegato o presentato in termini comprensibili a un pubblico particolare per uno scopo particolare.

Gli individui potrebbero cercare spiegazioni per diversi motivi. Avere una comprensione di come funziona un sistema potrebbe essere necessario esaminare e conoscere quanto bene un modello è funzionante; indagare le ragioni di un particolare risultato; o per gestire le interazioni sociali12. La natura e il tipo di spiegazione, trasparenza o giustificazione che richiedono varia in diversi contesti.

Questo briefing traccia la gamma di ragioni per cui diverse forme di spiegabilità potrebbero essere desiderabili per diversi individui o gruppi e le sfide che possono sorgere nel realizzarlo.

Il caso dell'IA spiegabile: come e perché l'interpretabilità è importante

C'è una serie di ragioni per cui potrebbe esserlo una qualche forma di interpretabilità nei sistemi di intelligenza artificiale auspicabile. Questi includono:

Dare agli utenti fiducia nel sistema: la fiducia degli utenti e la fiducia in un sistema di intelligenza artificiale sono spesso citate come ragioni per perseguire un'IA spiegabile. Le persone cercano spiegazioni per una varietà di scopi: supportare l'apprendimento, gestire le interazioni sociali, persuadere e assegnare responsabilità, tra gli altri13.

Tuttavia, la relazione tra l'affidabilità di un sistema e la sua spiegabilità non è semplice e l'uso di un'IA spiegabile per ottenere fiducia potrebbe dover essere trattato con cautela: ad esempio, spiegazioni apparenti plausibili potrebbero essere utilizzate per fuorviare gli utenti sul efficacia di un sistema14.

Protezione contro i pregiudizi: per verificare o confermare che un sistema di IA non utilizza i dati in modi che si traducono in pregiudizi o esiti discriminatori, è necessario un certo livello di trasparenza.

Soddisfare gli standard normativi oi requisiti delle politiche: la trasparenza o la spiegazione possono essere importanti per far valere i diritti legali che circondano un sistema, per dimostrare che un prodotto o servizio soddisfa gli standard normativi e per aiutare a risolvere le domande sulla responsabilità.
Esiste già una serie di strumenti politici che cercano di promuovere o imporre una qualche forma di spiegabilità nell'uso dei dati e dell'IA (descritta nell'allegato 1).

11. Si veda, ad esempio: Lipton, Z. (2016) The Mythos of Model Interpretability. Workshop ICML 2016 sull'interpretazione umana nell'apprendimento automatico (WHI 2016)

12. Miller, T. (2017). Spiegazione in Intelligenza Artificiale: Approfondimenti dalle Scienze Sociali. Intelligenza artificiale. 267. 10.1016/j.artt.2018.07.007.

13. Discusso più in dettaglio in Miller, T. (2017). Spiegazione in Intelligenza Artificiale: Approfondimenti dalle Scienze Sociali. Intelligenza artificiale. 267. 10.1016/j.artt.2018.07.007.

14. Vedi discussione successiva e Weller, A. (2017). Sfide per la trasparenza. Workshop sull'interpretazione umana (ICML 2017).

Miglioramento della progettazione del sistema:
l'interpretazione può consentire agli sviluppatori di interrogarsi sul motivo per cui un sistema si è comportato in un certo modo e sviluppare miglioramenti. Nelle auto a guida autonoma, ad esempio, è importante capire perché e come si è verificato un malfunzionamento di un sistema, anche se l'errore è solo di lieve entità. In ambito sanitario, l'interpretabilità può aiutare a spiegare risultati apparentemente anomali15.

Gli ingegneri progettano sistemi interpretabili per tenere traccia dei malfunzionamenti del sistema. I tipi di spiegazioni creati per svolgere questa funzione potrebbero assumere forme diverse da quelle richieste dai gruppi di utenti, sebbene entrambi possano includere l'analisi sia dei dati di addestramento che dell'algoritmo di apprendimento.

SCATOLA 2

Bias nei sistemi di intelligenza artificiale

I dati del mondo reale sono disordinati: contengono voci mancanti, possono essere distorti o soggetti a errori di campionamento e spesso vengono raccolti per scopi diversi dall'analisi in corso.

Errori di campionamento o altri problemi nella raccolta dei dati possono influenzare l'efficacia del il sistema di apprendimento automatico risultante funziona per utenti diversi. C'è stato un numero di istanze di alto profilo di sistemi di riconoscimento delle immagini che non funzionano in modo accurato per gli utenti di gruppi etnici minoritari, ad esempio.

I modelli creati da un sistema di machine learning possono anche generare problemi di equità o bias, anche se addestrati su dati accurati, e gli utenti devono essere consapevoli dei limiti dei sistemi che utilizzano. Nel reclutamento, ad esempio, i sistemi che fanno previsioni sugli esiti delle offerte di lavoro o della formazione possono essere influenzati da pregiudizi derivanti da strutture sociali che vi sono integrate dati al punto di raccolta. I modelli risultanti possono quindi rafforzare questi social pregiudizi, a meno che non vengano intraprese azioni correttive.

Concetti come l'equità possono avere significati diversi per comunità diverse e possono esserci dei compromessi tra questi interpretazioni diverse. Le domande su come costruire algoritmi "equi" sono oggetto di crescente interesse nelle comunità tecniche e idee su come farlo creare "correzioni" tecniche per affrontarli i problemi si stanno evolvendo, ma l'equità rimane una questione impegnativa. L'equità in genere implica l'applicazione dell'uguaglianza di alcune misure tra individui e/o gruppi, ma sono possibili molte nozioni diverse di equità: queste diverse nozioni possono spesso essere incompatibili, richiedendo più discussioni per negoziare inevitabili compromessi16.

15. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. e Elhadad, N. (2015) Modelli intelligibili per l'assistenza sanitaria: predizione del rischio di polmonite e riammissione ospedaliera a 30 giorni. KDD '15 Atti della 21a conferenza internazionale ACM SIGKDD sulla scoperta della conoscenza e l'estrazione di dati, 1721-1730

16. Kleinberg, J., Mullainathan, S. e Raghavan, M. (2016) Compromessi intrinseci nella determinazione equa dei punteggi di rischio, Atti della Conferenza internazionale ACM sulla misurazione e modellazione dei sistemi informatici, p40; e Kleinbery, J., Ludwig, J., Mullainathan, S. e Rambachan, A. (2018) Algorithmic fairness, Advances in Big Data Research in Economics, AEA Papers and Proceedings 2018, 108, 22-27

Valutare il rischio, la robustezza e la vulnerabilità:

Comprendere come funziona un sistema può essere importante nella valutazione del rischio17. Ciò può essere particolarmente importante se un sistema viene distribuito in un nuovo ambiente, in cui l'utente non può essere sicuro della sua efficacia. L'interpretazione può anche aiutare gli sviluppatori a capire in che modo un sistema potrebbe essere vulnerabile ai cosiddetti attacchi contraddittori, in cui gli attori che cercano di interrompere un sistema identificano un piccolo numero di punti dati scelti con cura da modificare per richiedere un output impreciso dal sistema.

Ciò può essere particolarmente importante nei compiti critici per la sicurezza18.

Comprendere e verificare gli output di un sistema: l' interpretazione può essere desiderabile nella verifica degli output di un sistema, tracciando come le scelte di modellazione, combinate con i dati utilizzati, influiscono sui risultati. In alcune applicazioni, questo può essere utile per aiutare gli sviluppatori a comprendere le relazioni di causa ed effetto nelle loro analisi19.

Autonomia, agenzia e rispetto dei valori sociali:

Per alcuni, la trasparenza è un valore sociale o costituzionale fondamentale e una parte fondamentale dei sistemi di responsabilità per attori potenti. Ciò riguarda le preoccupazioni relative alla dignità su come viene trattato un individuo in un processo decisionale. Una spiegazione può svolgere un ruolo nel sostenere l'autonomia individuale, consentendo a un individuo di contestare una decisione e aiutando a fornire un senso di azione nel modo in cui viene trattato20.

17. Nelle applicazioni finanziarie, ad esempio, gli investitori potrebbero non essere disposti a implementare un sistema senza comprenderne i rischi coinvolti o come potrebbe fallire, il che richiede un elemento di interpretabilità.

18. Vedi, ad esempio: S. Russell, D. Dewey e M. Tegmark (2015) "Research priority for robust and Benefit artificial intelligence", AI Magazine, vol. 36, n. 4, pp. 105–114.

19. Ad esempio, l'IA ha un'ampia gamma di applicazioni nella ricerca scientifica. In alcuni contesti, l'accuratezza da sola potrebbe essere sufficiente a rendere utile un sistema. Questo è discusso ulteriormente nel documento di discussione della Royal Society e dell'Alan Turing Institute sull'IA nella scienza, disponibile all'indirizzo: https://royalsociety.org/topics-policy/data-and-ai/artificial-intelligence/

20. Discusso in Burrell, J. (2016) How the Machine 'Thinks:' comprensione dell'opacità negli algoritmi di apprendimento automatico, Big Data & Society; e Ananny, M. & K. Crawford (2016). Vedere senza sapere: limiti dell'ideale di trasparenza e sua applicazione alla responsabilità algoritmica. Nuovi media e società. doi: https://doi.org/10.1177/1461444816676645

AI spiegabile: lo stato attuale delle cose

Esistono diversi approcci all'IA, che presentano diversi tipi di sfide di spiegabilità.

Gli approcci simbolici all'IA utilizzano tecniche basate su logica e inferenza. Questi approcci cercano di creare rappresentazioni simili a quelle umane dei problemi e l'uso della logica per affrontarli; i sistemi esperti, che funzionano a partire da set di dati che codificano la conoscenza e la pratica umana per automatizzare il processo decisionale, sono un esempio di tale approccio.

Sebbene l'IA simbolica in alcuni sensi si presti all'interpretazione – essendo possibile seguire i passaggi o la logica che hanno portato a un risultato – questi approcci incontrano ancora problemi di spiegabilità, con un certo livello di astrazione spesso richiesto per dare un senso al ragionamento su larga scala.

Gran parte del recente entusiasmo per i progressi in AI è il risultato dei progressi tecniche statistiche. Questi approcci, incluso l'apprendimento automatico, spesso sfruttano grandi quantità di dati e algoritmi complessi per identificare modelli e fare previsioni. Questa complessità, unita alla natura statistica delle relazioni tra gli input che il sistema costruisce, li rende difficili da comprendere, anche per utenti esperti, inclusi gli sviluppatori di sistema.

Riflettendo la diversità dei metodi di intelligenza artificiale che rientrano in queste due categorie, sono in fase di sviluppo molte diverse tecniche di intelligenza artificiale spiegabili. Questi si dividono – in linea di massima – in due gruppi:

- Lo sviluppo di metodi di intelligenza artificiale che sono intrinsecamente interpretabili, il che significa che la complessità o la progettazione del sistema è limitata per consentire a un utente umano per capire come funziona.
- L'uso di un secondo approccio che esamina come funziona il primo sistema di 'scatola nera', per fornire informazioni utili. Ciò include, ad esempio, metodi che rieseguono il modello iniziale con alcuni input modificati o che forniscono informazioni sull'importanza delle diverse funzionalità di input.

La tabella 1 fornisce una panoramica (non esaustiva) di alcuni di questi approcci. Questi forniscono diversi tipi di spiegazione, che includono: descrizioni del processo mediante il quale funziona un sistema; panoramiche del modo in cui un sistema crea una rappresentazione; e sistemi paralleli che generano un output e una spiegazione utilizzando modelli diversi.

Scelte effettuate nella selezione dei dati e nel modello il design influenza il tipo di spiegabilità che un sistema può supportare e approcci diversi hanno punti di forza e limiti diversi. Le mappe di salienza, ad esempio, possono aiutare un utente esperto a capire quali dati (o input) sono più rilevanti per il funzionamento di un modello, ma forniscono informazioni limitate su come tali informazioni vengono utilizzate21. Questo può essere sufficiente per alcuni scopi, ma rischia anche di tralasciare informazioni rilevanti.

21. Rudin, C. (2019) Smetti di spiegare i modelli di apprendimento automatico della scatola nera per decisioni ad alto rischio e utilizza interpretabili Modelli invece, Nature Machine Intelligence, 1, 206-215

TABELLA 1

Approcci diversi all'IA spiegabile affrontano diversi tipi di esigenze di spiegabilità e sollevano preoccupazioni diverse22.
Quali forme di spiegabilità dell'IA sono disponibili?

Che tipo di spiegazione si cerca	Quale metodo potrebbe essere appropriato?	Quali domande o preoccupazioni fanno questi i metodi aumentano?
Dettagli trasparenti di quale algoritmo viene utilizzato	**Pubblicazione dell'algoritmo**	Quale forma di spiegazione è più utile per coloro che sono interessati dall'esito del sistema? Il modulo di spiegazione fornito è accessibile alla comunità a cui è destinato? Quali processi di coinvolgimento degli stakeholder sono in atto per negoziare queste domande?
Come funziona il modello?	**Modelli intrinsecamente interpretabili** Utilizzare modelli la cui struttura e funzione è facilmente comprensibile per un utente umano, ad esempio un breve elenco di decisioni. **Sistemi scomponibili** Strutturare l'analisi in fasi, con un focus interpretabile su quei passaggi che sono più importanti nel processo decisionale. **Modelli proxy** Utilizzare un secondo modello, interpretabile, che corrisponde approssimativamente a un complesso sistema di "scatola nera".	Quali ulteriori controlli potrebbero essere necessari in altre fasi della pipeline decisionale? Ad esempio, come vengono fissati gli obiettivi del sistema? In che modo vengono utilizzati i diversi tipi di dati? Quali sono le implicazioni sociali più ampie dell'uso del sistema di intelligenza artificiale?
Quali input o caratteristiche dei dati sono più influenti nel determinare un output?	**Visualizzazione o mappatura della salienza** Illustrare in che misura le diverse funzioni di input influiscono sull'output di un sistema, in genere eseguito per un input di dati specifico.	Quanto è accurato e fedele il spiegazione fornita? C'è il rischio che possa fuorviare gli utenti? La forma di spiegazione desiderata è tecnicamente possibile in un dato contesto?
In un singolo caso, cosa dovrebbe cambiare per ottenere un output diverso?	**Spiegazioni controfattuali (o basate su esempi).** Generare spiegazioni focalizzate su un singolo caso, che identificano le caratteristiche dei dati di input che dovrebbero essere modificati per produrre un output alternativo.	

22. Adattato da Lipton, Z. (2016) The Mythos of Model Interpretability. Workshop ICML 2016 sull'interpretazione umana nell'apprendimento automatico (WHI 2016) e Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Spectre, M. e Kagal, L. (2018) Spiegazioni: an Panoramica sull'interpretabilità dell'apprendimento automatico. IEEE 5a conferenza internazionale sulla scienza dei dati. DOI:10.1109/dsaa.2018.00018

In questo contesto, la domanda per coloro che sviluppano e implementano l'IA non è semplicemente se sia spiegabile - o se un modello sia più spiegabile di un altro - ma se il sistema può fornire il tipo di spiegabilità necessario per un'attività specifica o un gruppo di utenti (laico o esperto, per esempio).

In considerazione di ciò, utenti e sviluppatori hanno esigenze diverse:

- Per gli utenti, spesso un approccio "locale", spiegare una decisione specifica, è molto utile. A volte, è importante consentire a un individuo di contestare un output, ad esempio contestare una domanda di prestito non andata a buon fine.

- Gli sviluppatori potrebbero aver bisogno di approcci "globali" che spieghino come funziona un sistema (ad esempio, per comprendere le situazioni in cui probabilmente funzionerà bene o male).

Gli approfondimenti della psicologia e delle scienze sociali indicano anche come i processi e i pregiudizi cognitivi umani possono influenzare l'efficacia di una spiegazione in diversi contesti:

- Gli individui tendono a cercare il contrasto spiegazioni – chiedendo perché una decisione è stata presa invece di un'altra – piuttosto che chiedersi solo perché si è verificato un determinato risultato;

- Le spiegazioni sono selettive, attingendo da a sottoinsieme dei fattori totali che hanno influenzato an risultato per spiegare perché è successo;

- Spiegazioni che fanno riferimento alle cause di un risultato sono spesso più accessibili o convincenti di quelli che fanno riferimento alle probabilità, anche se – nell'ambito dell'IA – il meccanismo è statistico piuttosto che causale;

- Il processo di spiegazione di qualcosa è spesso un'interazione sociale – uno scambio di informazioni tra due attori – che influenza il modo in cui vengono consegnati e ricevuti23.

I riquadri 3, 4 e 5 esplorano come alcuni di questi problemi si manifestano in contesti diversi.

23. Ciascuno di questi motivi di cui sopra è esplorato in dettaglio in Miller, T. (2017). Spiegazione in Intelligenza Artificiale: Approfondimenti da le scienze sociali. Intelligenza artificiale. 267. 10.1016/j.artt.2018.07.007.

RIQUADRO 3

Scienza

La raccolta e l'analisi dei dati è un elemento centrale del metodo scientifico e gli scienziati hanno a lungo utilizzato tecniche statistiche per aiutare il loro lavoro. All'inizio del 1900, ad esempio, lo sviluppo del t-test ha fornito ai ricercatori un nuovo strumento per estrarre informazioni dai dati al fine di verificare la veridicità delle loro ipotesi.

Oggi, l'apprendimento automatico è diventato uno strumento chiave per i ricercatori in tutti i domini per analizzare grandi set di dati, rilevare modelli precedentemente imprevisti o estrarre informazioni inaspettate. Le attuali aree di applicazione includono:

- Analisi dei dati genomici per prevedere le strutture proteiche, utilizzando approcci di apprendimento automatico in grado di prevedere la struttura tridimensionale delle proteine da sequenze di DNA;
- Comprendere gli effetti del clima cambiamento nelle città e nelle regioni, combinando dati di osservazione locali e modelli climatici su larga scala per fornire un quadro più dettagliato degli impatti locali dei cambiamenti climatici; e
- Trovare modelli nei dati astronomici, rilevamento di caratteristiche o segnali interessanti da grandi quantità di dati che potrebbero includere grandi quantità di rumore e classificazione queste caratteristiche per capire il diverso oggetti o modelli rilevati24.

In alcuni contesti, l'accuratezza di questi metodi da sola è sufficiente per rendere utile l'IA – filtrare le osservazioni del telescopio per identificare probabili bersagli per ulteriori studi, ad esempio. Tuttavia, l'obiettivo della scoperta scientifica è capire. I ricercatori vogliono sapere non solo qual è la risposta, ma perché.

L'IA spiegabile può aiutare i ricercatori a comprendere le informazioni che provengono dai dati della ricerca, fornendo interpretazioni accessibili di come i sistemi di IA conducono le loro analisi. Il progetto Automated Statistician, ad esempio, ha creato un sistema in grado di generare una spiegazione delle sue previsioni o previsioni, suddividendo complicati set di dati in sezioni interpretabili e spiegandone i risultati all'utente in un linguaggio accessibile25. Questo aiuta entrambi i ricercatori ad analizzare grandi quantità di dati e aiuta a migliorare la loro comprensione delle caratteristiche di tali dati.

24. Discusso ulteriormente in Royal Society e Alan Turing Institute (2019) La rivoluzione dell'IA nella scienza, disponibile su https://royalsociety.org/topics-policy/data-and-ai/artificial-intelligence/

25. Ulteriori dettagli sono disponibili su: https://www.automaticstatistician.com/about/

RIQUADRO 4

Giustizia criminale

Gli strumenti di valutazione del rischio della giustizia penale analizzano la relazione tra le caratteristiche di un individuo (dati demografici, precedenti di reati e così via) e la sua probabilità di commettere un reato o di essere riabilitato. Gli strumenti di valutazione del rischio hanno una lunga storia di utilizzo nella giustizia penale, spesso nel contesto di fare previsioni sul probabile comportamento futuro dei recidivi. Per alcuni, tali strumenti offrono la speranza di un sistema più equo, in cui i pregiudizi umani o le percezioni socialmente influenzate su chi è un "rischio" hanno meno probabilità di influenzare il modo in cui un individuo viene trattato dal sistema giudiziario26.

L'uso di strumenti di valutazione del rischio abilitati dall'IA offre quindi la possibilità di aumentare l'accuratezza e la coerenza di questi sistemi predittivi.

Tuttavia, l'opacità di tali strumenti ha sollevato preoccupazioni negli ultimi anni, in particolare in relazione all'equità e alla capacità di contestazione una decisione.

In alcune giurisdizioni, esiste già una legislazione contro l'uso di caratteristiche protette, come la razza o il genere, quando si prendono decisioni sulla probabilità di recidiva di un individuo. Queste funzionalità possono essere escluse dall'analisi in un sistema abilitato all'IA. Tuttavia, anche quando questi

caratteristiche sono escluse, la loro associazione con altre caratteristiche possono "infornare" l'ingiustizia nel sistema27; ad esempio, escludendo le informazioni sull'etnia, ma includendo i dati del codice postale che potrebbero essere correlati a distretti ad alta popolazione provenienti da comunità minoritarie.

Senza una qualche forma di trasparenza, può essere difficile valutare come potrebbero tali pregiudizi influenzare il punteggio di rischio di un individuo.

26. Walklate, S. (2019) Che aspetto avrebbe un sistema di giustizia giusta? In Dekeseredy, W. e Currie, E. (a cura di) Progressive Justice in an Age of Repression, Routledge, Londra.

27. Berk, R., Heidari, H., Jabbari, S., Kearns, M. e Roth, A. (2017) Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods and Research, pubblicato per la prima volta online 2 luglio 2018: http://journals.sagepub.com/doi/10.1177/0049124118782533

Negli Stati Uniti, ci sono già stati esempi di sistemi abilitati all'IA associati a esiti giudiziari sleali28 e di coloro che sono stati colpiti dai suoi risultati che hanno cercato di contestarne i risultati29. Nei dibattiti che sono seguiti, la mancanza di trasparenza sull'uso dell'IA – a causa delle protezioni della PI e dei segreti commerciali – è stata al centro. Ciò ha sollevato dubbi sul fatto che un algoritmo "scatola nera" violi il diritto a un giusto processo; quali disposizioni per la spiegabilità o altre forme di controllo pubblico sono necessarie quando si sviluppano strumenti di IA per la diffusione nei settori delle politiche pubbliche; su come strumenti di intelligenza artificiale più spiegabili potrebbero bilanciare il desiderio di trasparenza con il rischio di rivelare informazioni personali sensibili su un individuo30; e sui modi in cui gli strumenti tecnologici appaiono neutrali o autorevoli potrebbero influenzare indebitamente i loro utenti31. Queste sono aree importanti per più ricerca.

Le proposte per affrontare queste preoccupazioni hanno incluso:

- fornitura di ulteriori informazioni a giudici e coloro che lavorano nel sistema giudiziario per aiutarli a interpretare i risultati di un sistema e formazione aggiuntiva per tali individui;
- l'uso di stime di confidenza per aiutare gli utenti a interpretare i risultati di un sistema;
- sistemi per valutare, monitorare e controllare gli strumenti algoritmici32.

28. Rapporto Partnership on AI (2018) sugli strumenti algoritmici di valutazione del rischio nel sistema di giustizia penale statunitense. Disponibile su: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-us-criminal justice-system/

29. Ad esempio: Stato v. Loomis (Wis 2016). Ulteriori informazioni disponibili su: https://harvardlawreview.org/2017/03/state-v telaio/

30. Rapporto Partnership on AI (2018) sugli strumenti algoritmici di valutazione del rischio nel sistema di giustizia penale statunitense. Disponibile su: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-us-criminal justice-system/

31. Hannah-Moffat, K. (2018) Governance algoritmica del rischio: analisi dei big data, attivismo razziale e informativo nei dibattiti sulla giustizia penale', Criminologia teorica. https://doi.org/10.1177/1362480618763582

32. Rapporto Partnership on AI (2018) sugli strumenti algoritmici di valutazione del rischio nel sistema di giustizia penale statunitense. Disponibile su: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-us-criminal justice-system/

RIQUADRO 5

Salute

L'imaging medico è uno strumento importante per i medici nella diagnosi di una serie di malattie e nell'informare le decisioni sui percorsi di trattamento. Le immagini utilizzate in queste analisi, ad esempio le scansioni di campioni di tessuto, richiedono esperienza per essere analizzate e interpretate. Poiché l'uso di tale imaging aumenta in diversi domini medici, questa esperienza è sempre più richiesta.

L'uso dell'IA per analizzare i modelli nelle immagini mediche e per fare previsioni sulla probabile presenza o assenza di malattie è un'area di ricerca promettente. Per essere veramente utili in ambito clinico, tuttavia, questi sistemi di intelligenza artificiale dovranno funzionare bene nella pratica clinica e sia i medici che i pazienti potrebbero voler comprendere il ragionamento alla base di una decisione. Se i medici o i pazienti non sono in grado di capire perché un'IA ha fatto una previsione specifica, potrebbero sorgere dilemmi su quanta fiducia avere in quella sistema, soprattutto quando i trattamenti che ne conseguono possono avere effetti che alterano la vita33.

Un recente progetto di ricerca di DeepMind e del Moorfield's Eye Hospital indica nuovi metodi che possono consentire ai medici di migliorare comprendere i sistemi di intelligenza artificiale nel contesto dell'imaging medico. Questo progetto ha esaminato oltre 14.000 scansioni della retina, creando un sistema di intelligenza artificiale che ha analizzato queste immagini per rilevare la malattia della retina. Nonostante l'utilizzo di tecniche di deep learning che di solito verrebbero prese in considerazione 'scatola nera', i ricercatori hanno costruito il sistema in modo che gli utenti umani fossero in grado di capire perché aveva formulato una raccomandazione sulla presenza o l'assenza di malattia. Questa spiegabilità è stata integrata nel sistema rendendolo scomponibile.

Il sistema stesso è costituito da due reti neurali, ciascuna delle quali svolge funzioni diverse:

- Il primo analizza una scansione, utilizzando deep imparare a rilevare le caratteristiche nell'immagine che sono illustrative della presenza (o assenza) di una malattia, ad esempio emorragie nei tessuti. Questo crea una mappa delle caratteristiche nell'immagine.

- Il secondo analizza questa mappa, utilizzando le caratteristiche individuate dal primo per presentare una diagnosi ai clinici, presentando anche una percentuale per illustrare la fiducia nell'analisi.

All'interfaccia di questi due sistemi, i medici sono in grado di accedere a un intermedio rappresentazione che illustra quali aree di un'immagine potrebbero suggerire la presenza di malattie degli occhi. Questo può essere integrato nei flussi di lavoro clinici e interrogato da esperti umani che desiderano comprendere i modelli in una scansione e perché è stata formulata una raccomandazione, prima di confermare quale processo di trattamento è adatto. I medici restano quindi nel ciclo di fare una diagnosi e può lavorare con i pazienti per confermare i percorsi di trattamento34.

33. Castelvecchi, D. (2016) Possiamo aprire la scatola nera dell'IA? Natura, 538, 20-23, doi:10.1038/538020a

34. De Fauw, J., Ledsam, J., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King , D., Karthikesalingam, A., Hughes, C., Raine, R., Hughes, J., Sim, D., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees , G. , Back , T. , Village , P. , Suleyman , M. , Corebise , J. , Keane , P. , Ronneberger , O. , 2005 . (2018) Apprendimento profondo clinicamente applicabile per la diagnosi e il rinvio nella malattia della retina. Medicina della natura, 24, 1342 – 1350 https://doi.org/10.1038/s41591-018-0107-6

Sfide e considerazioni quando si implementa un'IA spiegabile

Mentre aumentare la spiegabilità dei sistemi di IA può essere vantaggioso per molte ragioni, ci sono anche sfide nell'implementazione di un'IA spiegabile. Questi includono quanto segue:

Utenti diversi richiedono forme di spiegazione diverse in contesti diversi

Contesti diversi danno origine a diverse esigenze di spiegabilità. Come notato in precedenza, gli sviluppatori di sistemi potrebbero richiedere dettagli tecnici su come funziona un sistema di intelligenza artificiale, mentre le autorità di regolamentazione potrebbero richiedere garanzie su come vengono elaborati i dati e coloro che sono soggetti a una decisione potrebbero voler capire quali fattori hanno portato a un output che li ha influenzati. Potrebbe quindi essere necessario spiegare una singola decisione o raccomandazione in più modi, riflettendo le esigenze di un pubblico diverso e le questioni in gioco in situazioni diverse.

Ciò può richiedere diversi tipi di contenuto; ad esempio, informazioni tecniche per uno sviluppatore; informazioni accessibili per un utente non esperto. Può anche porre diversi tipi di richiesta sulla progettazione del sistema in ogni fase del percorso di analisi.

Per capire come funziona un sistema, gli utenti potrebbero (variabilmente) desiderare di interrogare: quali dati è stato utilizzato dal sistema, la provenienza di tali dati e perché quei dati sono stati selezionati; come funziona il modello e quali fattori influenzano una decisione; perché è stato ottenuto un determinato output. Per capire quale tipo di spiegazione è necessaria, sono necessari un attento coinvolgimento degli stakeholder e una progettazione del sistema.

RIQUADRO 6

Cosa dicono i risultati degli esercizi di dialogo pubblico sulla necessità di un'IA spiegabile?

Le giurie dei cittadini nel 2018 hanno esaminato le opinioni pubbliche sulla spiegabilità nell'IA in diverse aree di applicazione. Questi sono stati commissionati dal Greater Manchester Patient Safety Translational Research Center e dal

Information Commissioner's Office (ICO), in collaborazione con l'Alan Turing Institute e facilitato dalle giurie dei cittadini cic e dal Jefferson Centre. Le giurie hanno deliberato sull'importanza di avere una spiegazione nei sistemi abilitati all'IA per: diagnosticare gli ictus; selezionare i CV per l'assunzione in azienda; abbinare donatori e riceventi di trapianto di rene; e selezionare i trasgressori per programmi di riabilitazione nel sistema di giustizia penale. In ogni scenario, ai partecipanti è stato chiesto di considerare l'importanza relativa di avere una spiegazione del funzionamento del sistema di intelligenza artificiale e l'accuratezza complessiva del sistema.

I risultati di queste giurie hanno mostrato che il contesto è importante quando gli individui valutano la necessità di una spiegazione. I partecipanti attribuiscono un'elevata priorità alla disponibilità delle spiegazioni in alcuni contesti, mentre in altri lo sono ha indicato che altri fattori – in questo caso, precisione – erano più importanti. Le discussioni nelle giurie hanno indicato che questa variazione era collegata a diversi motivi per volere una spiegazione: se fosse necessario contestare una decisione o fornire un feedback in modo che un individuo potesse cambiare il proprio comportamento, allora una spiegazione diventava più importante.

Questi risultati mostrano che gli individui fanno diversi compromessi quando si valuta se un sistema di intelligenza artificiale è affidabile. Suggeriscono che, in alcuni casi, potrebbe essere accettabile implementare un sistema a "scatola nera" se può essere verificato come più accurato del metodo spiegabile disponibile.

Ad esempio, nelle situazioni sanitarie, i giurati hanno indicato che avrebbero dato la priorità all'accuratezza di un sistema, mentre nella giustizia penale hanno attribuito un valore elevato all'avere una spiegazione per impugnare una decisione35.

I risultati della giuria dei cittadini dell'ICO fanno eco ai risultati dei dialoghi pubblici della Royal Society del 2016 e del 2017 sull'apprendimento automatico.

In questi dialoghi, la maggior parte delle persone non aveva sentito il termine "apprendimento automatico" - solo il 9% degli intervistati lo riconosceva - ma la maggioranza si era imbattuto almeno in alcune delle sue applicazioni nella vita quotidiana. Ad esempio, il 76% degli intervistati aveva sentito parlare di computer in grado di riconoscere il parlato e rispondere alle domande, come si trova negli assistenti personali virtuali disponibili su molti smartphone.

Gli atteggiamenti nei confronti dell'apprendimento automatico, positivi o negativi, dipendevano dalle circostanze in cui veniva utilizzato. La natura o la portata delle preoccupazioni del pubblico e la percezione delle potenziali opportunità erano legate alla domanda in esame.
Le opinioni delle persone su particolari applicazioni dell'apprendimento automatico erano spesso influenzate dalla loro percezione di chi stava sviluppando la tecnologia e di chi ne avrebbe beneficiato.

35. Il progetto ICO (2019) spiega: relazione intermedia. Disponibile su: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

La progettazione del sistema ha spesso bisogno di bilanciare le richieste concorrenti

Le tecnologie di intelligenza artificiale attingono da una gamma di metodi e approcci, ciascuno con vantaggi e limiti diversi. Il metodo di intelligenza artificiale utilizzato in qualsiasi applicazione influenzerà le prestazioni del sistema su più livelli, tra cui accuratezza, interpretabilità e privacy.

Precisione

Esistono diversi approcci alla creazione di sistemi interpretabili. Alcune IA sono interpretabili in base alla progettazione; questi tendono ad essere mantenuti relativamente semplici. Un problema con questi sistemi è che non possono ottenere la massima personalizzazione da grandi quantità di dati consentita da tecniche più complesse, come il deep learning. Ciò crea un compromesso tra prestazioni e precisione quando si utilizzano questi sistemi in alcune impostazioni, il che significa che potrebbero non essere desiderabili per quelle applicazioni in cui l'elevata precisione è apprezzata rispetto ad altri fattori. La natura di questi requisiti varierà a seconda dell'area di applicazione: i membri del pubblico hanno aspettative diverse sui sistemi utilizzati nell'assistenza sanitaria rispetto a quelli utilizzati

reclutamento, per esempio36.

Modelli diversi si adattano a compiti diversi - per alcuni problemi strutturati, ci possono essere poche differenze nelle prestazioni tra i modelli che sono intrinsecamente interpretabili e i sistemi "scatola nera". Altri problemi richiedono diverse forme di analisi dei dati, attingendo ai metodi della "scatola nera".

Privacy

In alcuni sistemi di IA – in particolare quelli che utilizzano dati personali o quelli in cui sono in gioco informazioni proprietarie – la richiesta di

la spiegazione può interagire con le preoccupazioni sulla privacy.

In aree come l'assistenza sanitaria e la finanza, ad esempio, un sistema di intelligenza artificiale potrebbe analizzare dati personali sensibili per prendere una decisione o una raccomandazione. Nel considerare il tipo di spiegabilità che potrebbe essere desiderabile in questi casi, le organizzazioni che utilizzano l'IA dovranno tenerne conto

quali diverse forme di trasparenza potrebbero comportare il rilascio di informazioni sensibili sugli individui37 o esporre potenzialmente a danni i gruppi vulnerabili38.

Potrebbero esserci anche casi in cui l'algoritmo oi dati su cui è stato addestrato sono proprietari, con le organizzazioni riluttanti a rivelare l'uno o l'altro per motivi di lavoro.

Ciò solleva interrogativi sul fatto che tali sistemi debbano essere implementati in aree in cui la comprensione del motivo per cui è stata presa una decisione è importante per la fiducia del pubblico o per garantire la responsabilità. Nelle recenti revisioni dell'applicazione dell'IA nella giustizia penale, ad esempio, ci sono state richieste di escludere l'uso di sistemi incomprensibili39.

36. Il progetto ICO (2019) spiega: relazione intermedia. Disponibile su: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

37. Weller, A. (2017) Sfide per la trasparenza, da Workshop on Human Interpretability in machine learning (WHI), ICML 2017.

38. Schudson M (2015) L'ascesa del diritto alla conoscenza. Cambridge, MA: Belknap Publishing.

39. Rapporto Partnership on AI (2018) sugli strumenti algoritmici di valutazione del rischio nel sistema di giustizia penale statunitense. Disponibile su: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-us-criminal justice-system/

La qualità e la provenienza dei dati fanno parte della pipeline di spiegabilità

Molti dei metodi di intelligenza artificiale di maggior successo odierni si basano sulla disponibilità di grandi quantità di dati per fare previsioni o fornire analisi per prendere decisioni informate. Comprendere la qualità e la provenienza dei dati utilizzati nei sistemi di IA è quindi una parte importante per garantire che un sistema sia spiegabile.

Coloro che lavorano con i dati devono capire come sono stati raccolti e le limitazioni a cui possono essere soggetti. Ad esempio, i dati utilizzati per creare sistemi di riconoscimento delle immagini potrebbero non funzionare bene per i gruppi minoritari (cfr. riquadro 2), i dati dei social media potrebbero riflettere solo determinate comunità di utenti o i dati dei sensori delle città potrebbero riflettere solo determinati tipi di quartiere. Spiegare quali dati sono stati utilizzati da un sistema di IA e come può quindi essere una parte importante per garantire che un sistema sia spiegabile.

La spiegazione può avere degli aspetti negativi

Le spiegazioni non sono necessariamente affidabili

Uno dei vantaggi proposti per aumentare la spiegabilità dei sistemi di intelligenza artificiale è una maggiore fiducia nel sistema: se gli utenti capiscono cosa ha portato a una decisione o raccomandazione generata dall'intelligenza artificiale, saranno più fiduciosi nei suoi risultati40.

Tuttavia, non solo il legame tra spiegazioni e fiducia è complesso41, ma la fiducia in un sistema potrebbe non essere sempre un risultato desiderabile42. Esiste il rischio che, se un sistema produce spiegazioni convincenti ma fuorvianti, gli utenti possano sviluppare un falso senso di fiducia o comprensione, credendo erroneamente che di conseguenza sia affidabile. Tale fiducia mal riposta potrebbe anche incoraggiare gli utenti a investire troppa fiducia nell'efficacia o nella sicurezza dei sistemi, senza che tale fiducia sia giustificata43. Le spiegazioni potrebbero aiutare ad aumentare la fiducia a breve termine, ma non aiutano necessariamente a creare sistemi che generano output affidabili o assicurano che coloro che implementano il sistema facciano affermazioni affidabili sulle sue capacità.

Possono esserci anche limitazioni all'affidabilità di alcuni approcci per spiegare perché un'IA ha raggiunto un determinato output. Nel caso della creazione di spiegazioni post-hoc, ad esempio, in cui un sistema di intelligenza artificiale viene utilizzato per analizzare i risultati di un altro sistema (non interpretabile), esiste la possibilità di generare spiegazioni plausibili: il secondo sistema sembra funzionare in modo identico a il primo – ma impreciso.

40. Miller, T. (2017). Spiegazione in Intelligenza Artificiale: Approfondimenti dalle Scienze Sociali. Intelligenza artificiale. 267. 10.1016/j.artt.2018.07.007.

41. Miller, T. (2017). Spiegazione in Intelligenza Artificiale: Approfondimenti dalle Scienze Sociali. Intelligenza artificiale. 267. 10.1016/j.artt.2018.07.007.

42. O'Neil, O. (2018) Collegare la fiducia all'affidabilità. Rivista internazionale di studi filosofici, 26, 2, 293 – 300 https://doi.org/10.1080/09672559.2018.1454637

43. Kroll, J. (2018) L'errore dell'imperscrutabilità, Phil. Trans. R. Soc. A 376: 20180084 https://doi.org/10.1098/rsta.2018.0084

L'utente ha una spiegazione su come funziona il sistema, ma che è staccata dal funzionamento effettivo dell'IA, sfruttando diverse caratteristiche dei dati o tralasciando aspetti importanti delle informazioni.

Un ulteriore rischio è l'inganno: l'uso di strumenti che generino interpretazioni plausibili che – intenzionalmente o meno – ingannano o manipolano le persone. Ad esempio, c'è una crescente pressione per aumentare la trasparenza sulla pubblicità mirata sui social media, in modo che gli utenti siano in grado di capire meglio come i dati su di loro vengono utilizzati dalle società di social media. In risposta, diverse aziende stanno sviluppando strumenti che spiegano ai propri utenti perché hanno visto contenuti particolari.
Tuttavia, non è chiaro se questi strumenti siano efficaci nell'aiutare gli utenti a capire come le loro interazioni online influenzino il contenuto vedono. Uno studio sulla motivazione data per il motivo per cui gli utenti hanno ricevuto determinati annunci ha rilevato che le spiegazioni fornite erano costantemente fuorvianti, mancando dettagli chiave che avrebbero consentito agli utenti di comprendere e potenzialmente influenzare il modo in cui sono stati presi di mira44.

Gioco

La trasparenza può svolgere un ruolo importante nel sostenere l'autonomia individuale: se gli utenti hanno accesso alle informazioni su come a è stata presa una decisione o una raccomandazione, potrebbero essere in grado di modificare il loro comportamento per ottenere un risultato più favorevole in futuro (discusso di seguito).

Tuttavia, questo legame tra trasparenza e capacità di influenzare gli output del sistema può avere conseguenze indesiderabili in alcuni contesti. Ad esempio, le autorità che indagano sui modelli di evasione fiscale possono cercare le caratteristiche di una dichiarazione dei redditi correlate all'evasione. Se questi indicatori sono ampiamente conosciuti, coloro che cercano di evadere le tasse potrebbero modificare il proprio comportamento per evitare di essere scoperti in modo più efficace45. Anche in questo caso, è probabile che la misura in cui si tratti di un problema varierà tra le applicazioni e sarà influenzata dall'obiettivo politico generale del sistema46.

La spiegazione da sola non può rispondere alle domande sulla responsabilità

La capacità di indagare e impugnare decisioni che hanno un impatto significativo su un individuo è fondamentale per i sistemi di responsabilità e l'obiettivo di alcuni approcci normativi attuali. In questo contesto, l'IA spiegabile può contribuire ai sistemi di responsabilità fornendo agli utenti l'accesso a informazioni e approfondimenti che consentono loro di impugnare una decisione o modificare il proprio comportamento per ottenere un risultato diverso in futuro.

Una varietà di fattori influenza la misura in cui un individuo è effettivamente in grado di contestare una decisione. Sebbene la spiegabilità possa essere una, anche le strutture organizzative, i processi di ricorso e altri fattori svolgeranno un ruolo nel plasmare il modo in cui un individuo può interagire con il sistema. Per creare un ambiente che supporti l'autonomia individuale e crei un sistema di responsabilità, è quindi probabile che siano necessari ulteriori passi.

Questi potrebbero includere, ad esempio, processi attraverso i quali gli individui possono contestare l'output di un sistema o contribuire a plasmarne la progettazione e il funzionamento47.

44. Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. e Weller, A. (2018) Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations, Proceedings of the 24th Network and Simposio sulla sicurezza dei sistemi distribuiti (NDSS), San Diego, California, febbraio 2018

45. Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D. e Yu, H. (2017) Accountable Algorithms, University of Pennsylvania Law Review, 165 , 633

46. Edwards L. e Veale M. (2017). Schiavo dell'algoritmo? Perché un "diritto a una spiegazione" probabilmente non è il rimedio per te Stanno cercando, Duke Law & Technology Review, 16(1), 18–84, doi:10.2139/ssrn.2972855.

47. Si veda, ad esempio: https://afog.berkeley.edu/files/2018/08/AFOG_workshop_panel3_report.pdf

Spiegare l'IA: dove dopo?

Il coinvolgimento degli stakeholder è importante per definire quale forma di spiegabilità è utile

Poiché le tecnologie di intelligenza artificiale vengono applicate su larga scala e in ambiti della vita in cui le conseguenze delle decisioni possono avere impatti significativi, aumenteranno le pressioni per sviluppare metodi di intelligenza artificiale i cui risultati possono essere compresi da diverse comunità di utenti. La ricerca sull'IA spiegabile sta avanzando, con l'emergere di una varietà di approcci. La misura in cui questi approcci sono utili dipenderà dalla natura e dal tipo di spiegazione richiesta.

Diversi tipi di spiegazioni saranno più o meno utili per diversi gruppi di persone che sviluppano, implementano, influenzano o regolano decisioni o previsioni dell'IA, e queste varieranno tra le aree di applicazione.

È improbabile che ci sia un metodo o una forma di spiegazione che funzioni su questi diversi gruppi di utenti. Le collaborazioni tra le discipline di ricerca e con i gruppi di stakeholder interessati da un sistema di intelligenza artificiale saranno importanti per aiutare a definire quale tipo di spiegazione è utile o necessaria in un determinato contesto e nella progettazione di sistemi per fornirle. Ciò richiede il lavoro oltre i confini della ricerca e dell'organizzazione per portare in primo piano prospettive o aspettative diverse prima che un sistema venga implementato.

La spiegazione potrebbe non essere sempre la priorità nella progettazione di un sistema di intelligenza artificiale; oppure potrebbe essere solo il punto di partenza

Esistono già esempi di sistemi di intelligenza artificiale che non sono facilmente spiegabili, ma possono essere implementati senza dare adito a preoccupazioni sulla spiegabilità: meccanismi di smistamento dei codici postali, ad esempio, o sistemi di raccomandazione negli acquisti online. Questi casi si trovano generalmente in aree in cui non ci sono conseguenze significative da risultati inaccettabili e l'accuratezza del sistema è stata ben convalidata48. Anche nelle zone dove il sistema potrebbe avere impatti significativi, se la qualità dei risultati è elevata, il sistema potrebbe comunque godere di elevati livelli di fiducia dell'utente49.

La necessità di spiegabilità deve essere considerata nel contesto del più ampio obiettivi o intenzioni per il sistema, tenendo conto delle domande sulla privacy, l'accuratezza degli output di un sistema, la sicurezza di un sistema e come potrebbe essere sfruttato da utenti malintenzionati se il suo funzionamento è noto e la misura in cui realizzare un sistema spiegabile potrebbe sollevare preoccupazioni sulla proprietà intellettuale o sulla privacy. Non si tratta di compromessi lineari – una maggiore spiegabilità che porta a una minore accuratezza, ad esempio – ma invece di progettare un sistema adatto alle esigenze che gli vengono poste.

48. Doshi-Velez F., Kim B. (2018) Considerazioni per la valutazione e la generalizzazione nell'apprendimento automatico interpretabile. In: Escalante H. et al. (a cura di) Modelli spiegabili e interpretabili in Computer Vision e Machine Learning. La serie Springer sulle sfide nell'apprendimento automatico. Springer, Cham

49. Cfr., ad esempio, indicazioni dal dialogo pubblico in: ICO (2019) Il progetto spiega: relazione intermedia. Disponibile su: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

In altri casi, la spiegabilità potrebbe essere una precondizione necessaria, ma che deve essere accompagnata da strutture più profonde per costruire la fiducia degli utenti o sistemi di responsabilità.

Se l'obiettivo desiderato è quello di responsabilizzare gli individui nelle loro interazioni con l'IA, ad esempio, allora potrebbero essere necessari meccanismi di feedback più ampi che consentano a coloro che interagiscono con il sistema di interrogarne i risultati, contestarli o essere in grado di alterare i propri risultati Altri significati. Quelli che progettano un sistema
dovrà quindi considerare come si adatta l'IA
nel più ampio contesto socio-tecnico della sua distribuzione. Data la gamma di metodi di intelligenza artificiale che esistono oggi, è anche vero che spesso esistono approcci meno complessi - le cui proprietà sono ben comprese - che possono dimostrare prestazioni forti come "black
metodi della scatola. Il metodo selezionato deve essere adatto alla sfida in corso.

Processi complessi spesso circondano il processo decisionale umano in domini critici e potrebbe essere necessario sviluppare un ambiente più ampio di responsabilità attorno all'uso dell'IA

Una linea argomentativa contro il dispiegamento di un'IA spiegabile punta alla mancanza di interpretabilità in molte decisioni umane: le azioni umane possono essere difficili da spiegare, quindi perché gli individui dovrebbero aspettarsi che l'IA sia diversa? Tuttavia, in molte aree critiche, tra cui l'assistenza sanitaria, la giustizia e altri servizi pubblici, nel tempo si sono sviluppati processi decisionali per mettere in atto procedure o salvaguardie per fornire diverse forme di responsabilità o audit. Le decisioni umane importanti spesso richiedono spiegazioni, conferimento o seconde opinioni e sono soggette a meccanismi di ricorso, audit e altre strutture di responsabilità. Questi riflettono complesse interazioni tra questioni tecniche, politiche, legali ed economiche, basandosi su modalità di controllo del funzionamento delle istituzioni che includono meccanismi esplicativi e non esplicativi50. La trasparenza e la spiegabilità dei metodi di IA possono quindi essere solo il primo passo nella creazione di sistemi affidabili.

50. Kroll, J. (2018) L'errore dell'imperscrutabilità, Phil. Trans. R. Soc. A 376: 20180084 https://doi.org/10.1098/rsta.2018.0084

Il rapporto della Royal Society Science as an open enterprise richiedeva un ambiente di apertura intelligente nella conduzione della scienza51. In un tale ambiente, le informazioni sarebbero:

- Accessibile: le informazioni dovrebbero essere localizzate in modo tale che possa essere facilmente trovato e in una forma che possa essere utilizzata.
- Valutabile: le informazioni dovrebbero essere conservate in uno stato in cui si possono esprimere giudizi sulla sua affidabilità. Ad esempio, i dati dovrebbero essere differenziati per diversi pubblico, o potrebbe essere necessario divulgare altre informazioni sul dati che potrebbero influenzare quelli di un individuo valutazione della sua affidabilità.
- Intellegibile: il pubblico deve essere in grado di esprimere un giudizio o una valutazione ciò che viene comunicato.
- Utilizzabile: le informazioni dovrebbero essere in a formato in cui altri possono usarlo, con dati di background appropriati forniti.

Questi criteri rimangono rilevanti per i dibattiti odierni sullo sviluppo di un'IA spiegabile. Ulteriori azioni per creare un ambiente di interpretabilità intelligente, compresi sia metodi di intelligenza artificiale spiegabili sia sistemi di responsabilità più ampi, potrebbero contribuire a un'attenta gestione delle tecnologie di intelligenza artificiale. Questi sistemi dovrebbero tenere conto dell'intera pipeline di sviluppo e implementazione dell'IA, che include la considerazione di come vengono fissati gli obiettivi per il sistema, come viene addestrato il modello, quali dati vengono utilizzati e le implicazioni per l'utente finale e la società.

Sarà necessario un ulteriore impegno tra responsabili politici, ricercatori, cittadini e coloro che implementano sistemi abilitati all'intelligenza artificiale per creare un tale ambiente di gestione.

51. Royal Society (2012) La scienza come impresa aperta. Disponibile su: https://royalsociety.org/topics-policy/projects/science-public-enterprise/#targetText=The%20final%20report%2C%20Science%20as,research%20that%20riflette%20pubblico%20valori.

Allegato 1: L'ambiente politico per un'IA spiegabile: uno schema

Le strutture politiche e di governance che circondano l'uso dei dati e l'IA sono costituite da una configurazione di strumenti legali e normativi, standard e norme di condotta professionali e comportamentali. Nel Regno Unito, queste strutture includono meccanismi normativi per disciplinare l'uso dei dati, quadri giuridici e politici emergenti che disciplinano l'applicazione dell'IA in determinati settori e una serie di codici di condotta che cercano di influenzare il modo in cui gli sviluppatori progettano i sistemi di intelligenza artificiale. Di seguito uno schizzo di questo ambiente.

Normativa e legge sulla protezione dei dati

Il regolamento generale sulla protezione dei dati (GDPR) dell'UE è il regolamento sulla protezione dei dati e sulla privacy che disciplina l'uso dei dati personali nei paesi dell'UE. Negli ultimi anni si è discusso molto sul fatto che questo regolamento, entrato in vigore nel Regno Unito nel 2018, contenga un "diritto a una spiegazione".

L'articolo 22 del Regolamento pone l'accento sul “processo decisionale individuale automatizzato, compresa la profilazione”. Si afferma che il soggetto cui si riferiscono i dati “ha il diritto di non essere sottoposto a una decisione basata unicamente su un trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo su di lui significativamente lo stesso”.

L'articolo prevede che, qualora tale trattamento sia necessario, l'interessato avrà “almeno il diritto di ottenere l'intervento umano […] per esprimere il proprio punto di vista e per impugnare la decisione”. L'articolo 15 del regolamento prevede inoltre che, in caso di ricorso al processo decisionale automatizzato, la persona cui si riferisce dovrebbe poter accedere a "informazioni significative sulla logica utilizzata", su richiesta, mentre gli articoli 13 e 14 richiedono che tali informazioni siano fornite in modo proattivo52 .

Il considerando 71 afferma inoltre che un soggetto soggetto a processi decisionali automatizzati “dovrebbe essere soggetto a garanzie adeguate, che dovrebbero comprendere informazioni specifiche all'interessato e il diritto di ottenere l'intervento umano, di esprimere il proprio punto di vista, di ottenere un spiegazione della decisione raggiunta dopo tale valutazione e di impugnare la decisione”53.

La guida che accompagna il testo del GDPR fornisce ulteriori approfondimenti sulla natura della spiegazione che potrebbe essere rilevante, ad esempio affermando: "il GDPR richiede al titolare del trattamento di fornire informazioni significative sulla logica coinvolta, non necessariamente una spiegazione complessa degli algoritmi utilizzati o la divulgazione dell'algoritmo completo. Le informazioni fornite dovrebbero, tuttavia, essere sufficientemente complete per consentire all'interessato di comprendere i motivi della decisione".

52. Parlamento europeo e Consiglio dell'Unione europea. 2016 Regolamento Generale UE sulla Protezione dei Dati – Articolo 22. Gazzetta Ufficiale dell'Unione Europea 59, L119/1–L119/149.

53. In questo contesto, la guida implica che il termine "logica" sia utilizzato per descrivere la logica alla base di un decisione, piuttosto che un processo di deduzione matematica formale, sebbene questo sia un punto di discussione sia nell'IA che nelle comunità legali. Questa guida non ha lo stesso peso legale del testo del GDPR. Gruppo di lavoro articolo 20 (2018) Linee guida sul processo decisionale individuale automatizzato e sulla profilazione ai fini del regolamento 2016/679, disponibili all'indirizzo http://ec.europa.eu/justice/data-protection/index_en.htm

Le prime interpretazioni di queste disposizioni suggerivano che costituissero un "diritto a una spiegazione" di come fosse stata presa una decisione abilitata all'IA. Tuttavia, i dibattiti da allora notato che:

- L'articolo 22 si applica solo a tipi limitati di processo decisionale automatizzato (ADM) – quello che è esclusivamente automatizzato e ha un effetto legale o significativo – e non copre tutte le decisioni assistite dall'IA.
- qualsiasi implicito "diritto a una spiegazione", in la relazione con l'articolo 22 non è giuridicamente vincolante54, sollevando interrogativi sulla natura e la portata di qualsiasi obbligo legale di fornire una spiegazione55.
- le disposizioni del GDPR si applicano solo ai casi di trattamento di dati personali.
- il GDPR prevede specifiche esenzioni da tale disposizione nei casi in cui una decisione automatizzata sia necessaria per un contratto; se la decisione è autorizzata dal diritto dello Stato membro; e se un individuo acconsente esplicitamente a una decisione. Sono inoltre previste protezioni aggiuntive in relazione ai segreti commerciali e alla protezione della PI56.

I dibattiti continuano sulla natura di spiegazioni richieste dal GDPR57 – e i diversi modi in cui le sue altre disposizioni potrebbero richiedere trasparenza da parte di coloro che trattano dati e utilizzano sistemi di IA oltre a quelli specificamente contemplati dall'articolo 22. Potrebbe essere necessaria un'ulteriore interpretazione da parte dei tribunali nazionali ed europei per comprendere meglio i limiti di tali disposizioni58.

Nel Regno Unito, l'Information Commissioner's Office (ICO) sta sviluppando linee guida per le organizzazioni, per supportarle nella creazione di processi o sistemi che possono aiutare a spiegare le decisioni relative all'IA alle persone interessate59. Nell'ambito di questo lavoro, l'ICO e The Alan Turing Institute hanno svolto una serie di giurie di cittadini in collaborazione con il National Institute for Health Research (NIHR) Greater Manchester Patient Safety Translational Research Center (GM PSRCC).
In un rapporto intermedio di queste giurie, l'ICO rileva tre approfondimenti chiave:

- l'importanza del contesto nel plasmare le opinioni delle persone;
- la necessità di una migliore istruzione; e
- le sfide dell'implementazione di un'IA60 spiegabile.

54. I considerando costituiscono una fonte di orientamento sulla legge, intesi ad aiutarne l'interpretazione.

55. Wachter, S., Mittelstadt, B. e Floridi, L. (2017) Perché un diritto alla spiegazione del processo decisionale automatizzato non esiste nel regolamento generale sulla protezione dei dati, diritto internazionale sulla privacy dei dati, 7, 2, 76 –99, https://doi.org/10.1093/idpl/ipx005

56. Edwards L. e Veale M. (2017). Schiavo dell'algoritmo? Perché un "diritto a una spiegazione" probabilmente non è il rimedio che stai cercando Duke Law & Technology Review, 16(1), 18–84, doi:10.2139/ssrn.2972855

57. Si veda, ad esempio: Kaminski, M. (2018) The Right to Explanation, Explained. Ricerca di studi legali sulla legge del Colorado Carta n. 18-24; Berkeley Technology Law Journal, vol. 34, n. 1, 2019 http://dx.doi.org/10.2139/ssrn.3196985; Edwards L. e Veale M. (2017). Schiavo dell'algoritmo? Perché un "diritto a una spiegazione" probabilmente non è il rimedio che stai cercando Duke Law & Technology Review, 16(1), 18–84, doi:10.2139/ssrn.2972855; e Wachter, S., Mittelstadt, B. e Floridi, L. (2017) Perché un diritto alla spiegazione del processo decisionale automatizzato non esiste nel regolamento generale sulla protezione dei dati, Legge internazionale sulla privacy dei dati, 7, 2, 76– 99, https://doi.org/10.1093/idpl/ipx005

58. Royal Society e British Academy (2017) Gestione e utilizzo dei dati: governance nel 21° secolo. Disponibile su https://royalsociety.org/topics-policy/data-and-ai/

59. Il progetto ICO (2019) spiega: relazione intermedia. Disponibile su: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

60. Il progetto ICO (2019) spiega: relazione intermedia. Disponibile su: https://ico.org.uk/media/2615039/project-explain-20190603.pdf

Approcci politici emergenti in tutti i settori

In diversi settori, gli organismi di regolamentazione stanno studiando se prodotti o servizi abilitati all'intelligenza artificiale possano sollevare dubbi sulla spiegabilità e se i framework esistenti siano sufficienti per gestire eventuali problemi che potrebbero derivare:

- Nel settore dei trasporti, l'Automated and Electric Vehicles Act (2018) del Regno Unito prevede la responsabilità degli assicuratori o dei proprietari di veicoli autonomi. Questo risponde ad alcune domande sull'eventuale necessità di una spiegazione per attribuire la responsabilità di un incidente61.
- L'Autorità di condotta finanziaria è commissionare ricerche per ottenere una migliore comprensione delle sfide di spiegabilità che sorgono quando si applica l'IA nel settore finanziario e collaborare con l'Organizzazione internazionale delle commissioni sui valori mobiliari per sviluppare un quadro per l'applicazione dell'IA etica nei servizi finanziari62.
- La regolamentazione dei medicinali e della sanità L'autorità sta lavorando con gli inglesi Standards Institute per esaminare la portata a cui i quadri di regolamentazione esistenti dispositivi medici sono in grado di affrontare il sfide poste dall'IA63 e se potrebbero essere necessari nuovi standard in settori quali la trasparenza e la spiegabilità per convalidare l'efficacia di tali dispositivi64.

Gli organismi internazionali stanno inoltre sviluppando standard tecnici per aree tra cui: governance dei dati per bambini e studenti e meccanismi di trasparenza e responsabilità che tale governance dovrebbe includere; pratiche di raccolta dei dati del datore di lavoro e trattamento trasparente ed etico dei dati sui dipendenti; e trasparenza nei sistemi decisionali macchina-macchina65.

Codici di condotta e design etico

In tutto il mondo, aziende, governi e istituti di ricerca hanno pubblicato principi per lo sviluppo e la diffusione dell'IA. Molti di questi includono richieste di forme di spiegabilità o trasparenza, spesso come mezzo per creare meccanismi di responsabilità che circondano l'IA e garantire che coloro che sono soggetti a decisioni supportate dai sistemi di IA dispongano delle informazioni di cui potrebbero aver bisogno per contestare tali decisioni. Un elenco (non esaustivo) di tali principi si trova nella tabella 2.

61. Reed C. 2018 Come dovremmo regolare l'intelligenza artificiale. Fil. Trans. R. Soc. A 376: 20170360. http://dx.doi.org/10.1098/rsta.2017.0360
62. Si veda, ad esempio: https://www.fca.org.uk/news/speeches/future-regulation-ai-consumer-good
63. Ulteriori dettagli sono disponibili su: https://content.yudu.com/web/43fqt/0A43ghs/IssueTwoMarch2019/html/index.html?page=20&origine=lettore
64. Ulteriori dettagli sono disponibili su: https://www.bsigroup.com/globalassets/localfiles/en-gb/about-bsi/nsb/innovation/mhra-ai-paper-2019.pdf
65. IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems. Maggiori dettagli su: https://standards.ieee.org/news/2017/ieee_p7004.html

TAVOLO 2

Spiegabilità nei principi dell'IA.

Autore	Principio	Dichiarazione
governo del Regno Unito – Scienza dei dati Quadro Etico	Rendi il tuo lavoro trasparente e essere responsabile	"Più complessi diventano gli strumenti della scienza dei dati, più difficile può essere capire o spiegare il processo decisionale. Questa è una questione critica da considerare quando si esegue la scienza dei dati o qualsiasi analisi nel governo. È essenziale che la politica del governo sia basata su prove interpretabili al fine di rendere conto di un risultato politico66".
Google - I nostri principi	Sii responsabile alla gente	"Progetteremo sistemi di intelligenza artificiale che forniscano opportunità appropriate per feedback, spiegazioni pertinenti e appello. Le nostre tecnologie di intelligenza artificiale saranno soggette alla direzione e al controllo umano appropriati67".
Microsoft – Il nostro approccio all'IA	Trasparenza	"Progettare l'intelligenza artificiale per essere affidabile richiede la creazione di soluzioni che riflettano principi etici che sono profondamente radicati in valori importanti e senza tempo. […] I sistemi di IA dovrebbero essere comprensibili68."
OCSE e G2069 – Principi dell'IA	Trasparenza e spiegabilità	"Gli attori dovrebbero impegnarsi alla trasparenza e alla divulgazione responsabile in merito ai sistemi di IA. A tal fine, dovrebbero fornire informazioni significative, adeguate al contesto e coerenti con lo stato dell'arte: io. promuovere una comprensione generale dei sistemi di IA, ii. rendere le parti interessate consapevoli delle loro interazioni con i sistemi di IA, anche sul posto di lavoro, iii. per consentire alle persone interessate da un sistema di intelligenza artificiale di comprendere il risultato e, iv. per consentire a coloro che sono colpiti negativamente da un sistema di IA di contestarne l'esito sulla base di un chiaro e informazioni di facile comprensione sui fattori e la logica che è servita come base per la previsione, la raccomandazione o la decisione70".
Accademia di Pechino di Artificiale Intelligenza – Principi dell'IA di Pechino	Sii etico	"La ricerca e lo sviluppo dell'IA dovrebbe adottare approcci di progettazione etica per rendere il sistema affidabile. Ciò può includere, ma non solo: rendere il sistema il più equo possibile, ridurre possibili discriminazioni e pregiudizi, migliorarne la trasparenza, la spiegazione e la prevedibilità e rendere il sistema più tracciabile, verificabile e responsabile71.
Alto livello dell'UE Gruppo di esperti su AI – Etica linee guida per un'IA affidabile	Trasparenza	"[I] modelli di business relativi a dati, sistemi e intelligenza artificiale dovrebbero essere trasparenti. I meccanismi di tracciabilità possono aiutare a raggiungere questo obiettivo. Inoltre, i sistemi di IA e le relative decisioni dovrebbero essere spiegati in modo adeguato alle parti interessate. Gli esseri umani devono essere consapevoli che stanno interagendo con un sistema di intelligenza artificiale e devono essere informati delle capacità e dei limiti del sistema72".

66. Cabinet Office Data Science Ethics Framework, disponibile all'indirizzo https://www.gov.uk/guidance/6-make-your-work-transparent-and-be-accountable

67. Principi dell'IA di Google, disponibili all'indirizzo: https://www.blog.google/technology/ai/ai-principles/

68. Principi di Microsoft AI, disponibili all'indirizzo https://www.microsoft.com/en-us/ai/our-approach-to-ai

69. Dichiarazione del G20 sull'IA, disponibile all'indirizzo: https://g20trade-digital.go.jp/dl/Ministerial_Statement_on_Trade_and_Digital_Economy.pdf

70. Principi dell'IA dell'OCSE, disponibili all'indirizzo: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

71. Principi dell'IA di Pechino, disponibili all'indirizzo: https://www.baai.ac.cn/blog/beijing-ai-principles

72. Principi dell'IA del gruppo ad alto livello dell'UE, disponibili all'indirizzo: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Allegato 2: Gruppo direttivo

Gruppo direttivo

I membri di un gruppo direttivo che ha sviluppato contenuti per questo briefing sono elencati di seguito.
I membri hanno agito a titolo individuale e non rappresentativo, contribuendo al progetto sulla base delle proprie competenze.

Gruppo direttivo
Alan Bundy CBE FRS FRSE FREng, Professore di Ragionamento Automatico, Università di Edimburgo
Jon Crowcroft FRS FREng, Professore Marconi di Sistemi di Comunicazione, Università di Cambridge
Zoubin Ghahramani FRS, professore di ingegneria dell'informazione, Università di Cambridge, e capo scienziato, Uber
Nancy Reid OC FRS FRSC, Canada Research Chair in Statistical Theory and Applications, Università di Toronto
Adrian Weller, Direttore del programma per l'IA, The Alan Turing Institute

Personale della Royal Society

Personale della Royal Society
Dott.ssa Natasha McCarthy, capo della politica, dati
Jessica Montgomery, consulente politico senior

Partecipanti al seminario

Oltre alla letteratura citata, questo briefing attinge dalle discussioni in un seminario tenuto con l'American Academy of Arts and Sciences nel 2018, tenutosi secondo la Chatham House Rule. La Società desidera esprimere i suoi ringraziamenti a tutti coloro che hanno presentato e partecipato a questo seminario.

La Royal Society è una Fellowship di autogoverno di molti degli scienziati più illustri del mondo provenienti da tutte le aree della scienza, dell'ingegneria e della medicina. Lo scopo fondamentale della Società, come lo è stato sin dalla sua fondazione nel 1660, è riconoscere, promuovere e sostenere l'eccellenza nella scienza e incoraggiare lo sviluppo e l'uso della scienza a beneficio dell'umanità.

Le priorità strategiche della Società sottolineano il suo impegno per la scienza della massima qualità, per la ricerca guidata dalla curiosità e per lo sviluppo e l'uso della scienza a beneficio della società. Queste priorità sono:

•Promuovere l'eccellenza nella scienza

•Supporto della collaborazione internazionale

•Dimostrare l'importanza della scienza a tutti

Per maggiori informazioni

La Società Reale
6 – 9 Terrazza di Carlton House
Londra SW1Y 5AG

T +44 20 7451 2500

E science.policy@royalsociety.org
Su royalsociety.org

Ente di beneficenza registrato n. 207043

ISBN: 978-1-78252-433-5
Rilasciato: novembre 2019 DES6051