Spiegare le decisioni prese con l'IA

ico.
Information Commissioner's Office

The Alan Turing Institute

introduzione 3
Parte 1 Le basi per spiegare l'IA 4
Parte 2: Spiegazione pratica dell'IA 6
Parte 3: Cosa significa spiegare l'IA per la tua organizzazione 8
Appendice 1: Esempio di costruzione e presentazione di una spiegazione di una diagnosi di cancro 10
Allegato 2: Tecniche algoritmiche 15
Allegato 3: Modelli supplementari 20
Allegato 4: Ulteriori letture 30
Allegato 5: Casi di assurance basata su argomenti 36

introduzione

Questa guida co-badge dell'ICO e dell'Alan Turing Institute mira a fornire alle organizzazioni consigli pratici per aiutare a spiegare i processi, i servizi e le decisioni fornite o assistite dall'IA alle persone interessate.

A prima vista

Sempre più spesso, le organizzazioni utilizzano l'intelligenza artificiale (AI) per supportare o prendere decisioni sugli individui. Se questo è qualcosa che fai, o qualcosa a cui stai pensando, questa guida è per te.

La guida si compone di tre parti. A seconda del tuo livello di esperienza e della composizione della tua organizzazione, alcune parti potrebbero essere più rilevanti per te rispetto ad altre.

Parte 1: Le basi per spiegare l'IA

Rivolta ai DPO e ai team di conformità, la prima parte definisce i concetti chiave e ne delinea una serie di diversi tipi di spiegazioni. Sarà rilevante per tutti i membri di personale coinvolto nella sviluppo di sistemi di IA.

Parte 2: Spiegazione pratica dell'IA

Rivolta ai team tecnici, la seconda parte ti aiuta con gli aspetti pratici per spiegare queste decisioni e fornire spiegazioni ai singoli.

Ciò sarà utile principalmente per i team tecnici della tua organizzazione, tuttavia anche il tuo DPO e il team di conformità lo troveranno utile.

Parte 3: Cosa significa spiegare l'IA per la tua organizzazione

Rivolta all'alta dirigenza, la terza parte esamina i vari ruoli, politiche, procedure e documentazione che è possibile mettere in atto per garantire che la propria organizzazione sia configurata per fornire spiegazioni significative alle persone interessate. Questo è principalmente rivolto al team dirigenziale senior della tua organizzazione, tuttavia anche il tuo DPO e il team di conformità lo troveranno utile.

Parte 1 Le basi per spiegare l'IA

A proposito di questa guida

Qual è lo scopo di questa guida?

Questa guida ha lo scopo di aiutare le organizzazioni a spiegare le decisioni prese dai sistemi di intelligenza artificiale (AI) alle persone interessate. Questa guida è suddivisa in tre parti:

- Parte 1 – Le basi per spiegare l'IA (questa parte)
- Parte 2 – Spiegazione pratica dell'IA
- Parte 3 – Cosa significa spiegare l'IA per la tua organizzazione

Questa parte della guida delinea:

- definizioni;
- requisiti legali per spiegare l'IA;
- vantaggi e rischi della spiegazione dell'IA;
- tipo spiegazione;
- fattori contestuali; e
- principi che stanno alla base del resto della guida.

Ci sono diversi motivi per spiegare l'IA, incluso il rispetto della legge e la realizzazione di vantaggi per la tua organizzazione e la società in generale. Chiarisce come applicare le disposizioni sulla protezione dei dati associate alla spiegazione delle decisioni sull'IA, oltre a evidenziare altri regimi legali pertinenti al di fuori del mandato dell'ICO.

Questa guida non è un codice di condotta legale ai sensi del Data Protection Act 2018 (DPA 2018). Al contrario, miriamo a fornire informazioni che ti aiutino a rispettare una serie di normative e a dimostrare le "migliori pratiche".

Come dovremmo usare questa guida?

Questa sezione introduttiva è per tutti i tipi di pubblico. Contiene concetti e definizioni che sono alla base del resto della guida.

I responsabili della protezione dei dati (DPO) e il team di conformità della tua organizzazione troveranno utile principalmente la sezione del quadro giuridico.

Anche i team tecnici e l'alta dirigenza potrebbero aver bisogno di una certa consapevolezza del quadro giuridico, nonché dei vantaggi e dei rischi della spiegazione dei sistemi di IA alle persone interessate dal loro utilizzo.

Troverai anche le sezioni "a colpo d'occhio" di questa guida in questo documento di riepilogo. Questo riunisce gli elementi fondamentali della guida in un unico posto e ne facilita la rapida ricerca.

Se gestisci una PMI che elabora dati personali utilizzando l'intelligenza artificiale e hai dei dubbi, vale la pena ricordare che puoi ottenere ulteriore supporto dall'hub web delle PMI dell'ICO.

Qual è lo stato di questa guida?

Questa guida è stata emessa in risposta all'impegno nell'accordo sul settore dell'IA del governo, ma non è un codice di condotta legale ai sensi del DPA 2018, né è intesa come una guida completa sulla conformità alla protezione dei dati.

Questa è una guida pratica che definisce le buone pratiche per spiegare le decisioni alle persone che sono state prese utilizzando i sistemi di IA che elaborano i dati personali.

Perché questa guida è dell'ICO e dell'Alan Turing Institute?

L'ICO è responsabile della supervisione della protezione dei dati nel Regno Unito e l'Alan Turing Institute (The Turing) è l'istituto nazionale del Regno Unito per la scienza dei dati e l'intelligenza artificiale.

Nell'ottobre 2017, la professoressa Dame Wendy Hall e Jérôme Pesenti hanno pubblicato la loro recensione indipendente sulla crescita dell'industria dell'IA nel Regno Unito. La seconda delle raccomandazioni del rapporto per supportare l'adozione dell'IA prevedeva che l'ICO e The Turing:

ÿ

"...sviluppare un quadro per spiegare i processi, i servizi e le decisioni forniti dall'IA, per migliorare la trasparenza e la responsabilità."

Nell'aprile 2018, il governo ha pubblicato il suo AI Sector Deal. L'accordo ha incaricato l'ICO e The Turing di:

ÿ

"...lavorare insieme per sviluppare una guida per aiutare a spiegare le decisioni relative all'IA".

Il rapporto indipendente e il Sector Deal fanno parte degli sforzi in corso compiuti dalle autorità di regolamentazione e dai governi nazionali e internazionali per affrontare le più ampie implicazioni della trasparenza e dell'equità nelle decisioni sull'IA che hanno un impatto sugli individui, sulle organizzazioni e sulla società in generale.

Parte 2: Spiegazione pratica dell'IA

A proposito di questa guida

Qual è lo scopo di questa guida?

Questa guida ti aiuta con gli aspetti pratici per spiegare le decisioni assistite dall'IA e fornire spiegazioni alle persone. Ti mostra come:

- seleziona la spiegazione appropriata per il tuo settore e caso d'uso;
- scegliere un modello adeguatamente spiegabile; e
- utilizzare determinati strumenti per estrarre spiegazioni da modelli meno interpretabili.

Come dovremmo usare questa guida?

Questa guida è principalmente per i team tecnici, tuttavia anche i DPO e i team di conformità la troveranno utile. Copre i passaggi che puoi intraprendere per spiegare le decisioni assistite dall'IA alle persone. Inizia con il modo in cui puoi scegliere quale tipo di spiegazione è più rilevante per il tuo caso d'uso e quali informazioni dovresti mettere insieme per ogni tipo di spiegazione. Per la maggior parte dei tipi di spiegazione, è possibile ricavare queste informazioni dalle decisioni e dalla documentazione di governance dell'organizzazione.

Tuttavia, data l'importanza centrale di comprendere la logica alla base del sistema di intelligenza artificiale per le spiegazioni assistite dall'IA, forniamo ai team tecnici una guida completa alla scelta di modelli adeguatamente interpretabili. Questo dipende dal caso d'uso. Indichiamo anche come utilizzare strumenti supplementari per estrarre elementi del funzionamento del modello in sistemi a "scatola nera". Infine, ti mostriamo come puoi fornire la tua spiegazione, contenente i tipi di spiegazione pertinenti che hai scelto, nel modo più utile per il destinatario della decisione.

Qual è lo stato di questa guida?

Questa guida è stata pubblicata in risposta all'impegno nell'accordo sul settore dell'intelligenza artificiale del governo, ma non è un codice di condotta legale ai sensi del Data Protection Act 2018 (DPA 2018) né è intesa come una guida completa sulla conformità alla protezione dei dati.

Questa è una guida pratica che definisce le buone pratiche per spiegare le decisioni alle persone che sono state prese utilizzando i sistemi di IA che elaborano i dati personali.

Perché questa guida è dell'ICO e dell'Alan Turing Institute?

L'ICO è responsabile della supervisione della protezione dei dati nel Regno Unito e l'Alan Turing Institute (The Turing) è l'istituto nazionale del Regno Unito per la scienza dei dati e l'intelligenza artificiale.

Nell'ottobre 2017, la professoressa Dame Wendy Hall e Jérôme Pesenti hanno pubblicato la loro recensione indipendente sulla crescita dell'industria dell'IA nel Regno Unito. La seconda delle raccomandazioni del rapporto per supportare l'adozione dell'IA prevedeva che l'ICO e The Turing:

ÿ

"...sviluppare un quadro per spiegare i processi, i servizi e le decisioni forniti dall'IA, per migliorare la trasparenza e la responsabilità."

Nell'aprile 2018, il governo ha pubblicato il suo AI Sector Deal. L'accordo ha incaricato l'ICO e The Turing di:

ÿ

"...lavorare insieme per sviluppare una guida per aiutare a spiegare le decisioni relative all'IA".

Il rapporto indipendente e il Sector Deal fanno parte degli sforzi in corso compiuti dalle autorità di regolamentazione e dai governi nazionali e internazionali per affrontare le più ampie implicazioni della trasparenza e dell'equità nelle decisioni sull'IA che hanno un impatto sugli individui, sulle organizzazioni e sulla società in generale.

Parte 3: Cosa significa spiegare l'IA per la tua organizzazione

A proposito di questa guida

Qual è lo scopo di questa guida?

Questa guida copre i vari ruoli, politiche, procedure e documentazione che puoi mettere in atto per garantire che la tua organizzazione sia configurata per fornire spiegazioni significative alle persone interessate.

Come dovremmo usare questa guida?

Questo è principalmente per i dirigenti senior della tua organizzazione. Offre un quadro generale dei ruoli che hanno un ruolo da svolgere nel fornire una spiegazione al destinatario della decisione, sia direttamente che come parte del processo decisionale.

Anche i responsabili della protezione dei dati (DPO) e i team di conformità, nonché i team tecnici possono trovare utile la sezione relativa alla documentazione.

Qual è lo stato di questa guida?

Questa guida è stata pubblicata in risposta all'impegno nell'accordo sul settore dell'intelligenza artificiale del governo, ma non è un codice di condotta legale ai sensi del Data Protection Act 2018 (DPA 2018) né è intesa come una guida completa sulla conformità alla protezione dei dati.

Questa è una guida pratica che definisce le buone pratiche per spiegare le decisioni alle persone che sono state prese utilizzando i sistemi di IA che elaborano i dati personali.

Perché questa guida è dell'ICO e dell'Alan Turing Institute?

L'ICO è responsabile della supervisione della protezione dei dati nel Regno Unito e l'Alan Turing Institute (The Turing) è l'istituto nazionale del Regno Unito per la scienza dei dati e l'intelligenza artificiale.

Nell'ottobre 2017, la professoressa Dame Wendy Hall e Jérôme Pesenti hanno pubblicato la loro recensione indipendente sulla crescita dell'industria dell'IA nel Regno Unito. La seconda delle raccomandazioni del rapporto per supportare l'adozione dell'IA prevedeva che l'ICO e The Turing:

ÿ

"...sviluppare un quadro per spiegare i processi, i servizi e le decisioni forniti dall'IA, per migliorare la trasparenza e la responsabilità."

Nell'aprile 2018, il governo ha pubblicato il suo AI Sector Deal. L'accordo ha incaricato l'ICO e The Turing di:

ÿ

"...lavorare insieme per sviluppare una guida per aiutare a spiegare le decisioni relative all'IA".

Il rapporto indipendente e il Sector Deal fanno parte degli sforzi in corso compiuti dalle autorità di regolamentazione e dai governi nazionali e internazionali per affrontare le più ampie implicazioni della trasparenza e dell'equità nelle decisioni sull'IA che hanno un impatto sugli individui, sulle organizzazioni e sulla società in generale.

Appendice 1: Esempio di costruzione e presentazione di una spiegazione di una diagnosi di cancro

Riunendo la nostra guida, l'esempio seguente mostra come un'organizzazione sanitaria potrebbe utilizzare i passaggi che abbiamo delineato per aiutarla a strutturare il processo di costruzione e presentare la propria spiegazione a un paziente affetto.

Compito 1: selezionare i tipi di spiegazione prioritaria considerando il dominio, il caso d'uso e l'impatto sugli individui

In primo luogo, l'organizzazione sanitaria familiarizza con i tipi di spiegazione in questa guida. Sulla base del contesto sanitario e dell'impatto della diagnosi di cancro sulla vita del paziente, l'organizzazione sanitaria seleziona i tipi di spiegazione che ritiene siano una priorità da fornire ai pazienti soggetti alle sue decisioni assistite dall'IA. Documenta la sua giustificazione per queste scelte:

Tipi di spiegazione prioritaria:

Razionale – Giustificare il ragionamento alla base del risultato del sistema di intelligenza artificiale per mantenere la responsabilità e utile per i pazienti se le tecniche di visualizzazione della spiegazione dell'IA sono disponibili per i non esperti...

Impatto – A causa della situazione ad alto impatto (vita/morte), è importante che i pazienti comprendano gli effetti e le fasi successive...

Responsabilità – È probabile che il pubblico non esperto voglia sapere con chi interrogare l'output del sistema di intelligenza artificiale...

Sicurezza e prestazioni - Data la complessità dei dati e del dominio, ciò può aiutare a rassicurare i pazienti sull'accuratezza, la sicurezza e l'affidabilità dell'output del sistema di intelligenza artificiale...

Altri tipi di spiegazione:

Dati: dettagli semplici sui dati di input, nonché sul set di dati di addestramento/convalida originali e su eventuali dati di convalida esterni

Equità - Poiché si tratta probabilmente di dati biofisici, al contrario di dati sociali o demografici, sorgeranno problemi di equità in aree come la rappresentatività dei dati e i pregiudizi di selezione, quindi potrebbe essere necessario fornire informazioni sugli sforzi di mitigazione dei pregiudizi pertinenti a questi ...

L'organizzazione sanitaria formalizza questi tipi di spiegazione nella parte pertinente della sua politica sulla governance dell'informazione:

Politica di governance dell'informazione

Uso dell'IA

Spiegare le decisioni dell'IA ai pazienti

Tipi di spiegazioni:

- Razionale
- Impatto
- Responsabilità
- Sicurezza e prestazioni
- Dati
- Equità

Compito 2: raccogliere e pre-elaborare i dati in modo consapevole delle spiegazioni

I dati utilizzati dall'organizzazione sanitaria incidono sulla sua valutazione dell'impatto e del rischio. L'organizzazione sanitaria quindi sceglie i dati con attenzione e considera l'impatto del pre-trattamento per garantire che siano in grado di fornire una spiegazione adeguata al destinatario della decisione.

Razionale

Informazioni su come i dati sono stati etichettati e come ciò mostra i motivi per classificare, ad esempio, determinate immagini come tumori.

Responsabilità

Informazioni su chi o quale parte dell'organizzazione sanitaria (o, se il sistema viene acquistato, chi o quale parte dell'organizzazione terza venditrice) è responsabile della raccolta e del pretrattamento dei dati del paziente. Essere trasparenti sul processo da un capo all'altro può aiutare l'organizzazione sanitaria a creare fiducia nell'uso dell'IA.

Dati

Informazioni sui dati che sono stati utilizzati, come sono stati raccolti e puliti e perché è stato scelto per addestrare il modello. Dettagli sulle misure adottate per garantire che i dati fossero accurati, coerenti, aggiornati, equilibrati e completi.

Sicurezza e prestazioni

Informazioni sulle metriche di prestazione selezionate del modello e su come, dati i dati disponibili inclusi nella formazione del modello, l'organizzazione sanitaria o il fornitore di terze parti hanno scelto le misure relative all'accuratezza che hanno fatto. Inoltre, informazioni sulle misure adottate per garantire che la preparazione e la preelaborazione dei dati garantiscano la robustezza e l'affidabilità del sistema in condizioni di runtime difficili, incerte o contraddittorie.

Compito 3: costruisci il tuo sistema per assicurarti di essere in grado di estrarre informazioni rilevanti per una gamma di tipi di spiegazione

L'organizzazione sanitaria, o un fornitore di terze parti, decide di utilizzare una rete neurale artificiale per sequenziare ed estrarre informazioni dalle immagini radiologiche. Sebbene questo modello sia in grado di predire l'esistenza e i tipi di tumori, il carattere ad alta dimensione della sua elaborazione lo rende opaco.

Il team di progettazione del modello ha scelto strumenti supplementari di "mappatura della salienza" e "mappatura dell'attivazione della classe".

per aiutarli a visualizzare le regioni critiche delle immagini che sono indicative di tumori maligni. Questi strumenti rendono visibili le aree problematiche evidenziando le regioni anormali. Tali immagini potenziate dalla mappatura consentono quindi a tecnici e radiologi di acquisire una comprensione più chiara delle basi cliniche della previsione del cancro del modello di intelligenza artificiale.

Ciò consente loro di garantire che l'output del modello supporti la pratica medica basata sull'evidenza e che i risultati del modello siano integrati in altre prove cliniche che sottoscrivono il giudizio professionale di questi tecnici e radiologi.

Compito 4: Traduci la logica dei risultati del tuo sistema in utilizzabili e facilmente comprensibili motivi

Il sistema di intelligenza artificiale utilizzato dall'ospedale per rilevare il cancro produce un risultato, che è una previsione che una particolare area su una risonanza magnetica contenga una crescita cancerosa. Questa previsione risulta essere una probabilità, con un particolare livello di confidenza, misurato in percentuale. Gli strumenti di mappatura supplementari forniscono successivamente al radiologo una rappresentazione visiva della regione cancerosa.

Il radiologo condivide queste informazioni con l'oncologo e altri medici dell'équipe medica insieme ad altre informazioni dettagliate sulle misure delle prestazioni del sistema e sui suoi livelli di certezza.

Per il paziente, l'oncologo o altri membri dell'équipe medica, quindi, mettono questo in un linguaggio, o in un altro formato, che il paziente può capire. Un modo in cui i medici scelgono di farlo è mostrare visivamente al paziente la scansione e strumenti di visualizzazione supplementari per aiutare a spiegare il risultato del modello.
Evidenziare le aree che il sistema di IA ha segnalato è un modo intuitivo per aiutare il paziente a capire cosa sta succedendo. I medici indicano anche quanta fiducia hanno nel risultato del sistema di intelligenza artificiale in base alle sue metriche di prestazioni e incertezza, nonché alla valutazione di altre prove cliniche rispetto a queste misure.

Attività 5: Prepara gli implementatori a distribuire il tuo sistema di intelligenza artificiale

Poiché il tecnico e l'oncologo utilizzano entrambi il sistema di intelligenza artificiale nel loro lavoro, l'ospedale decide che hanno bisogno di una formazione su come utilizzare il sistema.

La formazione dell'implementatore copre:

- come dovrebbero interpretare i risultati che genera il sistema di IA, in base alla comprensione di come è stato progettato e dei dati su cui è stato formato;
- come dovrebbero comprendere e soppesare le prestazioni e i limiti di certezza del sistema (cioè come visualizzare e interpretare le matrici di confusione, gli intervalli di confidenza, le barre di errore, ecc.);
- che dovrebbero utilizzare il risultato come una parte del loro processo decisionale, come complemento alla loro conoscenza del settore esistente;
- che dovrebbero esaminare criticamente se il risultato del sistema di IA si basa su una logica e una logica appropriate; e
- che in ogni caso dovrebbero preparare un piano per comunicare al paziente il risultato del sistema di IA e il ruolo che quel risultato ha giocato nel giudizio del medico.

Ciò include eventuali limitazioni nell'utilizzo del sistema.

Compito 6: Considera come costruire e presentare la tua spiegazione

I tipi di spiegazione scelti dall'organizzazione sanitaria hanno ciascuno una spiegazione basata sul processo e sull'esito. La qualità di ogni spiegazione è influenzata anche dal modo in cui raccolgono e preparano i dati di addestramento e test per il modello di intelligenza artificiale che scelgono. Raccolgono quindi le seguenti informazioni per ogni tipo di spiegazione:

Razionale

Spiegazione basata sul processo: informazioni per dimostrare che il sistema di IA è stato impostato in modo tale da consentire l'estrazione di spiegazioni della sua logica sottostante (direttamente o utilizzando strumenti supplementari); e che queste spiegazioni sono significative per i pazienti interessati.

Spiegazione basata sui risultati: informazioni sulla logica alla base dei risultati del modello e su come gli implementatori hanno incorporato tale logica nel loro processo decisionale. Ciò include il modo in cui il sistema trasforma i dati di input in output, come questi vengono tradotti in un linguaggio comprensibile ai pazienti e come il team medico utilizza i risultati del modello per raggiungere una diagnosi per un caso particolare.

Responsabilità

Spiegazione basata sul processo: informazioni sui responsabili all'interno delle organizzazioni sanitarie, o fornitori terzi, della gestione della progettazione e dell'uso del modello di IA e su come hanno assicurato che il modello fosse gestito in modo responsabile durante la sua progettazione e utilizzo.

Spiegazione basata sull'esito: informazioni sui responsabili dell'utilizzo dell'output del sistema di intelligenza artificiale come prova per supportare la diagnosi, per esaminarla e per fornire spiegazioni su come è nata la diagnosi (ovvero a chi può rivolgersi il paziente per interrogare la diagnosi).

Sicurezza e prestazioni

Spiegazione basata sul processo: informazioni sulle misure adottate per garantire la sicurezza generale e le prestazioni tecniche (sicurezza, accuratezza, affidabilità e robustezza) del modello di IA, comprese le informazioni sui test, la verifica e la convalida effettuati per certificarli.

Spiegazione basata sui risultati: informazioni sulla sicurezza e sulle prestazioni tecniche (sicurezza, accuratezza, affidabilità e robustezza) del modello di intelligenza artificiale nel suo funzionamento effettivo, ad esempio informazioni che confermano che il modello ha funzionato in modo sicuro e secondo la progettazione prevista nello specifico caso del paziente. Ciò potrebbe includere le misure di sicurezza e prestazioni utilizzate.

Impatto

Spiegazione basata sul processo: misure adottate durante la progettazione e l'utilizzo del modello di IA per garantire che non influisca negativamente sul benessere del paziente.

Spiegazione basata sugli esiti: informazioni sugli effettivi impatti del sistema di IA sul paziente.

L'organizzazione sanitaria considera quali fattori contestuali possono avere un effetto su ciò che i pazienti vogliono sapere sulle decisioni assistite dall'IA che intende prendere su una diagnosi di cancro. Stila un elenco dei fattori rilevanti:

Fattori contestuali

Dominio: test di sicurezza regolamentati...

Dati – biofisici…

Emergenza – se cancro, urgente...

Impatto: elevato, critico per la sicurezza...

Pubblico - per lo più non esperto...

L'organizzazione sanitaria sviluppa un modello per fornire la loro spiegazione delle decisioni dell'IA sulla diagnosi del cancro in modo stratificato:

Strato 1

- Spiegazione razionale
- Spiegazione dell'impatto
- Spiegazione di responsabilità
- Spiegazione di sicurezza e prestazioni

Consegna – es. il medico fornisce la spiegazione faccia a faccia con il paziente, integrata da informazioni cartacee/e-mail.

Livello 2

- Spiegazione dei dati
- Spiegazione di equità

Consegna: ad esempio, il medico fornisce al paziente queste informazioni aggiuntive in formato cartaceo/e-mail o tramite un'app.

Allegato 2: Tecniche algoritmiche

Algoritmo tipico	Descrizione di base	Possibili usi	Interpretabilità
Regressione lineare (LR)	Esegue previsioni su una variabile di destinazione sommando le variabili di input/ predittore ponderate.	Vantaggioso in settori altamente regolamentati come la finanza (es. credit scoring) e l'assistenza sanitaria (prevedere il rischio di malattia dato ad esempio lo stile di vita e le condizioni di salute esistenti) perché è più semplice da calcolare e controllare.	Elevato livello di interpretabilità a causa della linearità e della monotonia. Può diventare meno interpretabile con un numero maggiore di caratteristiche (ad es. dimensionalità elevata).
Regressione logistica	Estende la regressione lineare ai problemi di classificazione utilizzando una funzione logistica per trasformare gli output in una probabilità compresa tra 0 e 1.	Come la regressione lineare, vantaggiosa in settori altamente regolamentati e critici per la sicurezza, ma in casi d'uso basati su problemi di classificazione come decisioni sì/no su rischi, credito o malattia.	Buon livello di interpretabilità ma meno di LR perché le caratteristiche vengono trasformate attraverso una funzione logistica e correlate al risultato probabilistico in modo logaritmico piuttosto che come somme.
Regressione regolarizzata (LASSO e Cresta)	Estende la regressione lineare aggiungendo penalizzazione e regolarizzazione ai pesi delle caratteristiche per aumentare la scarsità/ridurre la dimensionalità.	Come la regressione lineare, vantaggiosa in settori altamente regolamentati e critici per la sicurezza che richiedono risultati comprensibili, accessibili e trasparenti.	Elevato livello di interpretabilità dovuto al miglioramento della scarsità del modello attraverso migliori procedure di selezione delle caratteristiche.
Modello lineare generalizzato (GLM)	Per modellare le relazioni tra caratteristiche e variabili target che non seguono distribuzioni normali (gaussiane), un GLM introduce una funzione di collegamento che consente l'estensione di LR a distribuzioni non normali.	Questa estensione di LR è applicabile ai casi d'uso in cui le variabili target hanno vincoli che richiedono l'insieme familiare esponenziale di distribuzioni (ad esempio, se una variabile target coinvolge numero di persone, unità di tempo o probabilità di esito, il risultato deve avere un valore non -valore negativo).	Buon livello di interpretabilità che tiene traccia dei vantaggi di LR introducendo anche una maggiore flessibilità. A causa della funzione di collegamento, determinare l'importanza della caratteristica può essere meno semplice rispetto al carattere additivo di un semplice LR, un certo grado di trasparenza potrebbe andare perso.

Modello additivo generalizzato (GAM)	Per modellare le relazioni non lineari tra le caratteristiche e le variabili target (non acquisite da LR), un GAM somma le funzioni non parametriche delle variabili predittive (come spline o adattamento basato su albero) piuttosto che semplici funzioni pesate.	Questa estensione di LR è applicabile ai casi d'uso in cui la relazione tra predittore e variabili di risposta non è lineare (cioè dove la relazione input-output cambia a velocità diverse in momenti diversi) ma si desidera un'interpretazione ottimale.	Buon livello di interpretabilità perché, anche in presenza di relazioni non lineari, la GAM consente una chiara rappresentazione grafica degli effetti delle variabili predittive sulle variabili di risposta.
albero decisionale (DT)	Un modello che utilizza metodi di ramificazione induttiva per suddividere i dati in nodi decisionali interconnessi che terminano in classificazioni o previsioni. DT si sposta dai nodi "radice" iniziali ai nodi "foglia" terminali, seguendo un percorso decisionale logico determinato da operatori "se-allora" di tipo booleano che vengono ponderati attraverso l'addestramento.	Poiché la logica passo-passo che produce risultati DT è facilmente comprensibile per gli utenti non tecnici (a seconda del numero di nodi/ caratteristiche), questo metodo può essere utilizzato in situazioni di supporto decisionale ad alto rischio e critiche per la sicurezza che richiedono trasparenza in quanto così come molti altri casi d'uso in cui il volume delle funzionalità rilevanti è ragionevolmente basso.	Elevato livello di interpretabilità se il DT viene mantenuto maneggevole, in modo che la logica possa essere seguita end-to-end. Il vantaggio di DT rispetto a LR è che il primo può ospitare non linearità e interazione variabile pur rimanendo interpretabile.
Elenchi e insiemi di regole/decisioni	Strettamente correlati ai DT, gli elenchi e gli insiemi di regole/decisioni applicano serie di istruzioni if-then alle funzionalità di input al fine di generare previsioni. Mentre gli elenchi di decisioni sono ordinati e restringono la logica dietro un output applicando regole "altrimenti", i set di decisioni mantengono le singole affermazioni se-allora non ordinate e ampiamente indipendenti, mentre le ponderano in modo che il voto delle regole possa avvenire nella generazione di previsioni.	Come per i DT, poiché la logica che produce elenchi e insiemi di regole è facilmente comprensibile per gli utenti non tecnici, questo metodo può essere utilizzato in situazioni di supporto decisionale ad alto rischio e critiche per la sicurezza che richiedono trasparenza, nonché in molti altri casi d'uso in cui il una giustificazione chiara e completamente trasparente dei risultati è una priorità.	Gli elenchi e gli insiemi di regole hanno uno dei più alti gradi di interpretabilità di tutte le tecniche algoritmiche con prestazioni ottimali e non opache. Tuttavia, condividono anche con DT la stessa possibilità che i gradi di comprensibilità vadano perduti man mano che gli elenchi di regole si allungano o le serie di regole diventano più grandi.
Ragionamento per casi (CBR)/	Utilizzando esempi tratti da precedenti conoscenze umane, CBR	CBR è applicabile in qualsiasi dominio basato sull'esperienza	CBR è interpretabile in base alla progettazione. Utilizza esempi tratti da

Prototipo e critica	prevede le etichette dei cluster apprendendo prototipi e organizzando le caratteristiche di input in sottospazi rappresentativi dei cluster di pertinenza. Questo metodo può essere esteso per utilizzare la discrepanza media massima (MMD) per identificare "critiche" o sezioni dello spazio di input in cui un modello rappresenta maggiormente i dati in modo errato. Una combinazione di prototipi e critiche può quindi essere utilizzata per creare modelli interpretabili in modo ottimale.	il ragionamento è utilizzato per il processo decisionale. Ad esempio, in medicina, i trattamenti sono raccomandati su base RBC quando precedenti successi in casi simili inducono il decisore a suggerire quel trattamento. L'estensione della CBR ai metodi di prototipazione e critica ha significato una migliore facilitazione della comprensione di complesse distribuzioni di dati e un aumento di insight, attuabilità e interpretabilità nel data mining.	conoscenza umana al fine di sottrarre caratteristiche di input in rappresentazioni umane riconoscibili. Preserva la spiegabilità del modello sia attraverso caratteristiche sparse che prototipi familiari.
Modello intero lineare supersparso (SLIM)	SLIM utilizza l'apprendimento basato sui dati per generare un semplice sistema di punteggio che richiede solo agli utenti di aggiungere, sottrarre e moltiplicare alcuni numeri per fare una previsione. Poiché SLIM produce un modello così sparso e accessibile, può essere implementato in modo rapido ed efficiente da utenti non tecnici, che non necessitano di una formazione speciale per implementare il sistema.	SLIM è stato utilizzato in applicazioni mediche che richiedono un processo decisionale clinico rapido e snello ma perfettamente accurato. Una versione denominata Risk-Calibrated SLIM (RiskSLIM) è stata applicata al settore della giustizia penale per dimostrare che i suoi metodi lineari sparsi sono efficaci per la previsione della recidiva quanto alcuni modelli opachi in uso.	A causa del suo carattere scarso e facilmente comprensibile, SLIM offre un'interpretazione ottimale per il supporto decisionale incentrato sull'uomo. In quanto sistema di punteggio completato manualmente, garantisce anche il coinvolgimento attivo dell'interprete-utente, che lo implementa.
Naïve Bayes	Usa la regola di Bayes per stimare la probabilità che una caratteristica appartenga a una data classe, assumendo che le caratteristiche siano indipendenti l'una dall'altra. Per classificare una caratteristica, il classificatore Naïve Bayes calcola la probabilità a posteriori per la classe	Sebbene questa tecnica sia chiamata ingenua a causa dell'assunto irrealistico dell'indipendenza delle caratteristiche, è nota per essere molto efficace. Il suo rapido tempo di calcolo e la sua scalabilità lo rendono adatto per applicazioni con elevate dimensioni	I classificatori Naïve Bayes sono altamente interpretabili, perché la probabilità di appartenenza alla classe di ciascuna caratteristica è calcolata in modo indipendente. L'ipotesi che le probabilità condizionate delle variabili indipendenti siano statisticamente indipendenti, tuttavia, lo è

	appartenenza a quella caratteristica moltiplicando la probabilità a priori della classe con la probabilità condizionata di classe della caratteristica.	spazi caratteristici. Le applicazioni comuni includono il filtro antispam, i sistemi di raccomandazione e l'analisi del sentiment.	anche un punto debole, perché le interazioni tra le funzionalità non vengono considerate.
K-vicino più vicino (KNN)	Utilizzata per raggruppare i dati in cluster ai fini della classificazione o della previsione, questa tecnica identifica un quartiere di vicini più vicini attorno a un punto dati di interesse e trova il risultato medio di essi per la previsione o la classe più comune tra loro per la classificazione.	KNN è una tecnica semplice, intuitiva e versatile che ha ampie applicazioni ma funziona meglio con set di dati più piccoli. Poiché non è parametrico (non fa ipotesi sulla distribuzione dei dati sottostante), è efficace per i dati non lineari senza perdere l'interpretabilità. Le applicazioni comuni includono i sistemi di raccomandazione, il riconoscimento delle immagini e la valutazione e l'ordinamento dei clienti.	KNN parte dal presupposto che classi o risultati possono essere previsti osservando la vicinanza dei punti dati da cui dipendono ai punti dati che hanno prodotto classi e risultati simili. Questa intuizione sull'importanza della vicinanza/prossimità è la spiegazione di tutti i risultati KNN. Tale spiegazione è più convincente quando lo spazio delle funzionalità rimane piccolo, in modo che la somiglianza tra le istanze rimanga accessibile.
Supporta le macchine vettoriali (SVM)	Utilizza un tipo speciale di funzione di mappatura per creare un divisore tra due insiemi di caratteristiche in uno spazio di caratteristiche ad alta dimensione. Un SVM quindi ordina due classi massimizzando il margine del confine di decisione tra di loro.	Gli SVM sono estremamente versatili per compiti di smistamento complessi. Possono essere utilizzati per rilevare la presenza di oggetti nelle immagini (faccia/senza faccia; gatto/senza gatto), per classificare tipi di testo (articolo sportivo/articoli d'arte) e per identificare geni di interesse per la bioinformatica.	Basso livello di interpretabilità che dipende dalla dimensionalità dello spazio delle caratteristiche. Nei casi determinati dal contesto, l'uso di SVM dovrebbe essere integrato da strumenti esplicativi secondari.
Rete neurale artificiale (ANN)	Famiglia di tecniche statistiche non lineari (incluse reti neurali ricorrenti, convoluzionali e profonde) che costruiscono complesse funzioni di mappatura per prevedere o classificare i dati utilizzando il feedforward (e talvolta il feedback) delle variabili di input tramite	Le ANN sono più adatte per completare un'ampia gamma di attività di classificazione e previsione per spazi di caratteristiche ad alta dimensione, ad esempio casi in cui sono presenti vettori di input molto grandi. I loro usi possono variare dalla visione artificiale, al riconoscimento di immagini, alle vendite e alle previsioni meteorologiche,	Le tendenze alla curvatura (estrema non linearità) e all'elevata dimensionalità delle variabili di input producono livelli di interpretabilità molto bassi nelle RNA. Sono considerati l'epitome delle tecniche della "scatola nera". Ove appropriato, dovrebbe essere l'uso di ANN

	reti addestrate di operazioni interconnesse e multistrato.	scoperta farmaceutica e previsione delle scorte alla traduzione automatica, diagnosi di malattie e rilevamento di frodi.	integrato da strumenti esplicativi secondari.
A caso foresta	Crea un modello predittivo combinando e calcolando la media dei risultati di più (a volte migliaia) di alberi decisionali addestrati su sottoinsiemi casuali di funzionalità condivise e dati di addestramento.	Le foreste casuali vengono spesso utilizzate per aumentare efficacemente le prestazioni dei singoli alberi decisionali, per migliorare i loro tassi di errore e per mitigare l'overfitting. Sono molto popolari in aree problematiche ad alta dimensione come la medicina genomica e sono state anche ampiamente utilizzate nella linguistica computazionale, nell'econometria e nella modellazione predittiva del rischio.	Livelli di interpretabilità molto bassi possono derivare dal metodo di addestramento di questi insiemi di alberi decisionali su dati insacchettati e caratteristiche randomizzate, sul numero di alberi in una determinata foresta e sulla possibilità che i singoli alberi possano avere centinaia o addirittura migliaia di nodi.
Imposta i metodi	Come suggerisce il nome, i metodi d'insieme sono una classe diversificata di meta-tecniche che combina diversi modelli "studenti" (dello stesso o tipo diverso) in un modello più grande (predittivo o classificatorio) al fine di diminuire la distorsione statistica, diminuire la varianza o migliorare le prestazioni di uno qualsiasi dei sottomodelli presi separatamente.	I metodi Ensemble hanno un'ampia gamma di applicazioni che tengono traccia dei potenziali usi dei loro modelli costitutivi di apprendimento (questi possono includere DT, KNN, Random Forests, Naïve Bayes, ecc.).	L'interpretabilità dei metodi Ensemble varia a seconda del tipo di metodi utilizzati. Ad esempio, può essere difficile spiegare la logica di un modello che utilizza tecniche di bagging, che mediano insieme più stime di studenti formati su sottoinsiemi casuali di dati. Le esigenze esplicative di questo tipo di tecniche dovrebbero essere valutate caso per caso.

Allegato 3: Modelli supplementari

Strategia esplicativa supplementare	Che cos'è e a cosa serve?	Restrizioni
Modelli surrogati (SM)	Gli SM costruiscono un modello interpretabile più semplice (spesso un albero decisionale o un elenco di regole) dal set di dati e dalle previsioni di un sistema opaco. Lo scopo del SM è fornire una proxy comprensibile del modello complesso che stima bene quel modello, pur non avendo lo stesso grado di opacità. Sono utili per assistere nei processi di diagnosi e miglioramento del modello e possono aiutare a esporre l'overfitting e il bias. Possono anche rappresentare alcune non linearità e interazioni che esistono nel modello originale.	Come approssimazione, gli SM spesso non riescono a cogliere l'intera portata delle relazioni non lineari e delle interazioni ad alta dimensione tra le caratteristiche. C'è un compromesso apparentemente inevitabile tra la necessità che il SM sia sufficientemente semplice in modo che sia comprensibile dagli esseri umani e la necessità che quel modello sia sufficientemente complesso in modo da poter rappresentare la complessità di come la funzione di mappatura di un Il modello "scatola nera" funziona nel suo insieme. Detto questo, la misurazione R2 può fornire una buona metrica quantitativa dell'accuratezza dell'approssimazione del SM del modello complesso originale.
Globale/locale? Interna/post-hoc?	Per la maggior parte, gli SM possono essere utilizzati sia a livello globale che locale. In quanto proxy semplificati, sono post-hoc.	
Dipendenza parziale Trama (PDP)	Un PDP calcola e rappresenta graficamente l'effetto marginale di una o due caratteristiche di input sull'output di un modello opaco, sondando la relazione di dipendenza tra le variabili di input di interesse e il risultato previsto nel set di dati, calcolando la media dell'effetto di tutte le altre caratteristiche del modello. Questo è un buon strumento di visualizzazione, che consente una rappresentazione chiara e intuitiva del comportamento non lineare per funzioni complesse (come foreste casuali e SVM). È utile, ad esempio, per mostrare che un dato modello di interesse soddisfa i vincoli di monotonicità nella distribuzione a cui si adatta.	Mentre i PDP consentono preziosi accesso a relazioni non lineari tra predittore e variabili di risposta, e quindi anche per il confronto del comportamento del modello con le aspettative informate sul dominio di relazioni ragionevoli tra caratteristiche e risultati, non tengono conto delle interazioni tra le variabili di input in esame. Possono, in questo modo, essere fuorvianti quando alcune caratteristiche di interesse sono fortemente correlate con altre caratteristiche del modello. Poiché gli effetti marginali della media di PDP, possono anche essere fuorvianti se le caratteristiche hanno effetti irregolari la funzione di risposta su diversi sottoinsiemi di dati, ad es dove hanno diversi

		associazioni con l'output in punti diversi. Il PDP può appiattire queste eterogeneità alla media.
Globale/locale? Interna/post-hoc?	I PDP sono spiegazioni post-hoc globali che possono anche consentire una comprensione causale più profonda del comportamento di un modello opaco attraverso la visualizzazione. Queste intuizioni sono, tuttavia, molto parziali e incomplete sia perché i PDP non sono in grado di rappresentare interazioni di caratteristiche ed effetti eterogenei, sia perché non sono in grado di rappresentare graficamente più di un paio di caratteristiche alla volta (il pensiero spaziale umano è limitato a poche dimensioni , quindi solo due variabili nello spazio 3D sono facilmente comprensibili).	
individuale condizionale Trama delle aspettative (GHIACCIO)	Perfezionando ed estendendo i PDP, ICE traccia graficamente la relazione funzionale tra una singola caratteristica e la risposta prevista per una singola istanza. Mantenendo costanti tutte le caratteristiche tranne la caratteristica di interesse, i grafici ICE rappresentano come, per ogni osservazione, una determinata previsione cambia al variare dei valori di quella caratteristica. Significativamente, i grafici ICE quindi disaggregano o scompongono la media degli effetti delle caratteristiche parziali generati in un PDP mostrando i cambiamenti nella relazione di output delle caratteristiche per ogni istanza specifica, cioè l'osservazione per osservazione. Ciò significa che può sia rilevare le interazioni che tenere conto di associazioni irregolari di predittore e variabili di risposta.	Se usato in combinazione con I grafici PDP, ICE possono fornire informazioni locali sul comportamento delle funzionali che migliora le spiegazioni globali più grossolane offerte dai PDP. Soprattutto, i grafici ICE sono in grado di rilevare gli effetti di interazione e eterogeneità nelle caratteristiche che rimangono nascoste ai PDP in virtù del modo in cui calcolano la dipendenza parziale degli output dalle caratteristiche di interesse calcolando la media dell'effetto delle altre variabili predittive. Tuttavia, sebbene i grafici ICE possano identificare le interazioni, è anche probabile che manchino correlazioni significative tra le caratteristiche e diventare fuorviante in alcuni casi. Anche la costruzione di grafici ICE può diventare difficile quando i set di dati sono molto grandi. In questi casi, possono essere impiegate tecniche di approssimazione che consentono di risparmiare tempo, come l'osservazione del campionamento o il binning delle variabili (ma, a seconda delle regolazioni e delle dimensioni del set di dati, con un impatto inevitabile sull'accuratezza della spiegazione).

Globale/locale? Interna/post-hoc?	I grafici ICE offrono una forma locale e post-hoc di spiegazione supplementare.	
Locale accumulato Trame degli effetti (ALE)	Come approccio alternativo a I grafici PDP e ALE forniscono una visualizzazione dell'influenza di caratteristiche individuali sul previsioni di un modello "scatola nera" calcolando la media della somma delle differenze di previsione per istanze di caratteristiche di interesse ad intervalli localizzati e poi integrando questi effetti medi in tutti gli intervalli. In questo modo, sono in grado di rappresentare graficamente gli effetti locali accumulati delle caratteristiche sul funzione di risposta nel suo insieme. Poiché i grafici ALE utilizzano differenze locali nella previsione quando calcolano l'influenza media della caratteristica (invece del suo effetto marginale come fanno i PDP), è in grado di tenere conto meglio delle interazioni delle caratteristiche ed evitare distorsioni statistiche. Questo capacità di stimare e rappresentare l'influenza delle caratteristiche in a modo consapevole della correlazione è un vantaggio delle trame ALE. I grafici ALE sono anche più trattabili dal punto di vista computazionale di PDP perché sono in grado di utilizzare tecniche per calcolare gli effetti a intervalli più piccoli e frammenti di osservazioni.	Una notevole limitazione dei grafici ALE ha a che fare con il modo in cui suddividono la distribuzione dei dati in intervalli che sono in gran parte scelti dal progettista delle spiegazioni. Se ci sono troppi intervalli, le differenze di previsione possono diventare troppo piccole e le influenze di stima meno stabili. Se gli intervalli lo sono troppo allargato, il grafico cesserà di rappresentare sufficientemente la complessità del modello sottostante. Sebbene i grafici ALE siano utili per fornire spiegazioni globali che tengano conto delle correlazioni delle caratteristiche, dovrebbero essere considerati anche i punti di forza dell'utilizzo di PDP in combinazione con i grafici ICE (soprattutto quando ci sono meno effetti di interazione in il modello che viene spiegato). Tutte e tre le tecniche di visualizzazione fanno luce su diverse dimensioni di interesse nella spiegazione dei sistemi opachi, quindi l'adeguatezza del loro impiego dovrebbe essere valutata caso per caso.
Globale/locale? Interna/post-hoc?	I grafici ALE sono una forma globale e post-hoc di spiegazione supplementare.	
Variabile globale Importanza	La strategia di importanza della variabile globale calcola il contributo di ciascuna funzione di input per modellare l'output nel set di dati permutando il	Sebbene la permutazione delle variabili per misurare la loro importanza relativa, in una certa misura, tenga conto degli effetti di interazione, c'è ancora a

	caratteristica di interesse e misurazione dei cambiamenti nell'errore di previsione: se la modifica del valore della caratteristica permutata aumenta l'errore del modello, allora quella caratteristica è considerata importante. L'utilizzo dell'importanza della variabile globale per comprendere l'influenza relativa delle caratteristiche sulle prestazioni del modello può fornire informazioni significative sulla logica alla base del comportamento del modello. Questo metodo fornisce anche una preziosa comprensione delle non linearità nel modello complesso che viene spiegato.	alto grado di imprecisione nel metodo rispetto a quali variabili interagiscono e quanto sono queste interazioni incidendo sulle prestazioni del modello. Una più ampia limitazione del quadro dell'importanza della variabile globale deriva da ciò che è noto come "effetto Rashomon". Questo si riferisce alla varietà di diversi modelli che possono adattarsi ugualmente bene alla stessa distribuzione dei dati. Questi modelli possono avere insiemi molto diversi di caratteristiche significative. Poiché la tecnica basata sulla permutazione può fornire solo informazioni esplicative riguardo alle prestazioni di un singolo modello, non è in grado di affrontare questo più ampio problema della varietà di schemi esplicativi efficaci.
Globale/locale? Interna/post-hoc?	L'importanza della variabile globale è una forma di spiegazione globale e post-hoc.	
Variabile globale Interazione	La strategia di interazione delle variabili globali calcola l'importanza delle interazioni delle variabili nel set di dati misurando la varianza nella previsione del modello quando si presume che le variabili potenzialmente interagenti siano indipendenti. Ciò viene fatto principalmente calcolando una "statistica H" in cui una funzione di dipendenza parziale senza interazione viene sottratta da una funzione di dipendenza parziale osservata per calcolare la varianza nella previsione. Questa è una strategia esplicativa versatile, che è stata impiegata per calcolare gli effetti di interazione in molti tipi di modelli complessi tra cui ANN e Random Forests. Può essere usato per calcolare	Sebbene la capacità di base di identificare gli effetti di interazione in modelli complessi sia un contributo positivo dell'interazione variabile globale come strategia esplicativa supplementare, ci sono un paio di potenziali inconvenienti a cui potresti voler prestare attenzione. Innanzitutto, non esiste una metrica stabilita in questo metodo per determinare il soglia quantitativa attraverso la quale diventano le interazioni misurate significativo. Il significato relativo delle interazioni è un'informazione utile in quanto tale, ma non c'è modo di sapere a che punto le interazioni siano abbastanza forti da esercitare effetti.

	interazioni tra due o più variabili e anche tra variabili e la funzione di risposta nel suo insieme. È stato utilizzato efficacemente, ad esempio, nella ricerca biologica per identificare gli effetti di interazione tra i geni.	In secondo luogo, il carico computazionale di questa strategia esplicativa è molto elevato, poiché gli effetti di interazione vengono calcolati in modo combinatorio su tutti i punti dati. Ciò significa che all'aumentare del numero di punti dati, il numero di calcoli necessari aumenta in modo esponenziale.
Globale/locale? **Interna/post-hoc?**	L'interazione variabile globale è una forma di spiegazione globale e post-hoc.	
Analisi di sensibilità e Layer-Wise Rilevanza Propagazione (LRP)	L'analisi della sensibilità e l'LRP sono strumenti esplicativi supplementari utilizzati per le reti neurali artificiali. L'analisi della sensibilità identifica le caratteristiche più rilevanti di un vettore di input calcolando i gradienti locali per determinare come spostare un punto dati per modificare l'etichetta di output. Qui, la sensibilità di un output a tali cambiamenti nei valori di input identifica le caratteristiche più rilevanti. LRP è un altro metodo per identificare la rilevanza delle caratteristiche che è a valle dell'analisi di sensibilità. Utilizza una strategia di spostamento all'indietro attraverso gli strati di un grafico a rete neurale per mappare modelli di alta attivazione nei nodi e, in definitiva, genera raggruppamenti interpretabili di variabili di input salienti che possono essere rappresentate visivamente in una mappa di attribuzione di calore o pixel.	Sia l'analisi della sensibilità che l'LRP identificano variabili importanti negli ampi spazi delle caratteristiche delle reti neurali. Queste tecniche esplicative trovano schemi visivamente informativi mettendo insieme matematicamente i valori dei singoli nodi nel Rete. Come conseguenza di questo approccio frammentario, offrono ben poco come spiegazione del ragionamento o della logica alla base dei risultati dell'elaborazione dei dati di una RNA. Di recente, sempre più ricerche si sono concentrate sull'attenzione metodi per identificare le rappresentazioni di ordine superiore che guidano le funzioni di mappatura di questo tipo di modelli, nonché su metodi CBR interpretabili che sono integrati nelle architetture ANN e che analizzano le immagini identificando parti prototipiche e combinandole in un insieme rappresentativo. Queste nuove tecniche stanno dimostrando che sono stati compiuti alcuni progressi significativi nello scoprire la logica sottostante di alcune ANN.

Globale/locale? **Interna/post-hoc?**	L'analisi della sensibilità e la mappatura della salienza sono forme di spiegazione locale e post-hoc, sebbene la recente incorporazione delle tecniche di RBC stia spostando le spiegazioni della rete neurale verso una base di interpretazione più interna.	
Interpretabile locale Agnostico dal modello Spiegazione (LIME) e ancore	LIME funziona adattando un modello interpretabile a una specifica previsione o classificazione prodotta da un sistema opaco. Lo fa campionando punti dati casualmente attorno alla previsione o classificazione del target e quindi utilizzandoli per costruire un'approssimazione locale del confine decisionale che possa spiegare le caratteristiche che figurano in primo piano nella specifica previsione o classificazione sotto esame. LIME lo fa generando un semplice modello di regressione lineare pesando i valori dei punti dati, che sono stati prodotti perturbando casualmente il modello opaco, in base alla loro vicinanza alla previsione o classificazione originale. Il più vicino di questi valori all'istanza che vengono spiegati sono i più pesanti, in modo che il modello supplementare possa produrre una spiegazione dell'importanza della caratteristica che è localmente fedele a quell'istanza. Si noti che possono essere utilizzati anche altri modelli interpretativi come gli alberi decisionali.	Mentre LIME sembra essere un passo nella giusta direzione, nella sua versatilità e nella disponibilità di molte iterazioni in software molto utilizzabili, una serie di problemi che presentano sfide all'approccio rimangono irrisolti. Ad esempio, l'aspetto cruciale di come definire correttamente la misura di prossimità per il "quartiere" o la "regione locale" in cui si applica la spiegazione rimane poco chiaro e piccoli cambiamenti nella scala della misura scelta possono portare a spiegazioni molto divergenti. Allo stesso modo, la spiegazione prodotta dal modello lineare supplementare può diventare rapidamente inaffidabile, anche con perturbazioni piccole e praticamente impercettibili del sistema che sta tentando di approssimare. Ciò sfida l'assunto di base che esiste sempre un modello interpretabile semplificato che approssima con successo il modello sottostante ragionevolmente bene vicino a un dato punto dati. I creatori di LIME hanno ampiamente riconosciuto queste carenze e hanno recentemente offerto un nuovo approccio esplicativo che chiamano "ancora". Queste "regole di alta precisione" incorporano nelle loro strutture formali "modelli ragionevoli" che operano all'interno del modello sottostante (come le convenzioni linguistiche implicite che sono all'opera in un

		modello di previsione del sentimento), in modo che possano stabilire confini adeguati e fedeli della loro copertura esplicativa delle sue previsioni o classificazioni.
Globale/locale? Interna/post-hoc?	LIME offre una forma di spiegazione supplementare locale e post-hoc.	
Additivo Shapley Spiegazioni (SHAP)	SHAP utilizza concetti della teoria dei giochi cooperativi per definire un "valore di Shapley" per una caratteristica di interesse che fornisce una misurazione della sua influenza sulla previsione del modello sottostante. In generale, questo valore viene calcolato calcolando la media del contributo marginale della caratteristica a ogni possibile previsione per l'istanza in esame. Il modo in cui SHAP calcola i contributi marginali consiste nel costruire due istanze: la prima istanza include la funzione misurato, mentre il secondo lo esclude sostituendolo con una variabile sostitutiva selezionata casualmente. Dopo aver calcolato la previsione per ciascuna di queste istanze inserendo i loro valori nel modello originale, il risultato della seconda viene sottratto da quello di il primo a determinare il contributo marginale della caratteristica. Questa procedura viene quindi ripetuta per tutte le possibili combinazioni di funzionalità in modo che la media ponderata di tutti i contributi marginali della caratteristica di interesse può essere	Tra i numerosi inconvenienti di SHAP, il più pratico è che tale procedura è computazionalmente onerosa e diventa intrattabile oltre una certa soglia. Nota, tuttavia, alcune versioni successive di SHAP offrono metodi di approssimazione come Kernel SHAP e Shapley Sampling Values per evitare questa eccessiva spesa computazionale. Questi metodi, tuttavia, influiscono sull'accuratezza complessiva del metodo. Un'altra significativa limitazione di SHAP è che il suo metodo di campionamento dei valori al fine di misurare i contributi variabili marginali assume indipendenza delle caratteristiche (cioè che i valori campionati non sono correlati in modi che potrebbero influenzare in modo significativo l'output per un determinato calcolo). Di conseguenza, gli effetti di interazione generati da e tra le variabili sostitutive che vengono utilizzati come sostituti per le funzioni tralasciate sono necessariamente disattesi quando condizionali i contributi sono approssimativi. Il risultato è l'introduzione di incertezza nella spiegazione che viene prodotta, perché il

	calcolato. Questo metodo consente quindi a SHAP, per estensione, di stimare i valori di Shapley per tutte le funzionalità di input nel set per produrre la distribuzione completa della previsione per l'istanza. Sebbene computazionalmente intensivo, ciò significa che per il calcolo dell'istanza specifica, SHAP può garantire assiomaticamente la coerenza e l'accuratezza del suo calcolo dell'effetto marginale della caratteristica. Questo la robustezza computazionale ha reso SHAP attraente come an esplicativo per un'ampia varietà di modelli complessi, perché può fornire un quadro più completo dell'influenza relativa delle caratteristiche per una determinata istanza rispetto a qualsiasi altro strumento di spiegazione post-hoc.	la complessità delle interazioni multivariate nel modello sottostante potrebbe non essere sufficientemente catturata dalla semplicità di questa tecnica di interpretabilità supplementare. Questo inconveniente nel campionamento (così come un certo grado di arbitrarietà nella definizione del dominio) può rendere SHAP inaffidabile anche con perturbazioni minime del modello che sta approssimando. Attualmente sono in corso sforzi per tenere conto della funzionalità dipendenze nei calcoli SHAP. I creatori originali della tecnica hanno introdotto Tree SHAP per includere, almeno in parte, le interazioni delle funzionalità. Altri hanno estensioni introdotte di recente di Kernel SHAP.
Globale/locale? Interna/post-hoc?	SHAP offre una forma di spiegazione supplementare locale e post-hoc.	
Controfattuale Spiegazione	Le spiegazioni controfattuali offrono informazioni su come fattori specifici che hanno influenzato un la decisione algoritmica può essere modificata in modo che le migliori alternative possano essere realizzate dal destinatario di una particolare decisione o risultato. L'incorporazione di spiegazioni controfattuali in un modello al momento della consegna consente alle parti interessate di vedere quali variabili di input del modello possono essere modificato, in modo che il risultato potrebbero essere modificati a loro vantaggio. Per i sistemi di intelligenza artificiale che aiutano	Sebbene la spiegazione controfattuale offra un modo utile per esplorare in modo contrastante come l'importanza delle caratteristiche possa influenzare un risultato, presenta limitazioni che hanno origine nella varietà di possibili caratteristiche che possono essere incluse quando si considerano risultati alternativi. Incerto casi, il solo numero di caratteristiche potenzialmente significative che potrebbero essere in gioco nelle spiegazioni controfattuali di un dato risultato può rendere difficile ottenere una spiegazione chiara e diretta e serie selezionate di possibili spiegazioni sembrano potenzialmente arbitrarie.

	decisioni su azioni umane mutevoli (come decisioni di prestito o punteggio di credito), incorporando spiegazioni controfattuali nelle fasi di sviluppo e test dello sviluppo del modello può consentire l'incorporazione di variabili attuabili, ovvero variabili di input che forniranno ai soggetti decisionali opzioni concise per apportare modifiche pratiche che migliorerebbero le loro possibilità di ottenere il risultato desiderato. In questo modo, le strategie esplicative controfattuali possono essere utilizzate come modo per incorporare la ragionevolezza e il incoraggiamento dell'agenzia nella progettazione e implementazione di sistemi di IA.	Inoltre, ci sono ancora limitazioni sui tipi di set di dati e funzioni a cui sono applicabili questo tipo di spiegazioni. Infine, poiché questo tipo di spiegazione ammette apertamente l'opacità del modello algoritmico, è meno in grado di affrontare le preoccupazioni su interazioni delle caratteristiche potenzialmente dannose e relazioni di covariate discutibili che possono essere sepolte in profondità nell'architettura del modello. È una buona idea utilizzare spiegazioni controfattuali insieme ad altre strategie esplicative supplementari, ovvero come una componente di un portafoglio di spiegazioni più completo.
Globale/locale? Interna/post-hoc?	Le spiegazioni controfattuali sono una forma locale e post-hoc di strategia di spiegazione supplementare.	
Autoesplicativo e Basato sulla cautela Sistemi	I sistemi autoesplicativi e basati sull'attenzione integrano effettivamente strumenti esplicativi secondari nei sistemi opachi in modo che possano offrire spiegazioni runtime dei propri comportamenti. Ad esempio, un sistema di riconoscimento delle immagini potrebbe avere un componente primario, come una rete neurale convoluzionale, che estrae caratteristiche dai suoi input e li classifica mentre un componente secondario, come una rete neurale ricorrente incorporata con un meccanismo di "direzione dell'attenzione" traduce le caratteristiche estratte in un naturale rappresentazione linguistica che	L'automazione delle spiegazioni attraverso sistemi autoesplicativi è un approccio promettente per le applicazioni in cui gli utenti traggono vantaggio dall'acquisizione di informazioni in tempo reale sulla logica dei sistemi complessi che stanno utilizzando. Tuttavia, indipendentemente dalla loro utilità pratica, questi tipi di strumenti secondari funzioneranno solo così come l'infrastruttura esplicativa che sta effettivamente decomprimendo le loro logiche sottostanti. Questo livello esplicativo deve rimanere accessibile ai valutatori umani e essere comprensibile agli interessati individui. I sistemi autoesplicativi, in altre parole, dovrebbero rimanere essi stessi interpretabili in modo ottimale. Tè

	fornisce una spiegazione lunga frase del risultato al uso. La ricerca sull'integrazione di interfacce "basate sull'attenzione". continuando a progredire verso il potenziale rendere le loro implementazioni più sensibili alle esigenze degli utenti, spiegazioni future e umanamente comprensibili. Inoltre, l'incorporazione della conoscenza del dominio e delle strutture basate sulla logica o sulle convenzioni le architetture di modelli complessi consentono sempre più di incorporare rappresentazioni e prototipi migliori e più facili da usare.	il compito di formulare una strategia primaria di spiegazione supplementare è ancora parte del processo di costruzione di un sistema con capacità di autoesplicazione. Un'altra potenziale trappola da considerare per i sistemi autoesplicativi è la loro capacità di fuorviare o di fornire false rassicurazioni agli utenti, specialmente quando le qualità umane sono incorporate nel loro metodo di consegna. Ciò può essere evitato non progettando qualità antropomorfe nella loro interfaccia utente e rendendo esplicite le metriche di incertezza ed errore nella spiegazione man mano che viene fornita.
Globale/locale? Interna/post-hoc?	Poiché i sistemi autoesplicativi e basati sull'attenzione sono strumenti secondari che possono utilizzare molti metodi diversi di spiegazione, possono essere globali o locali, interni o post-hoc o una combinazione di uno qualsiasi di essi.	

Allegato 4: Ulteriori letture

Fonte standard PROV

Moreau, L. & Missier, P. (2013). *PROV-DM: Il modello di dati PROV.* Raccomandazione W3C.
URL : https://www.w3.org/TR/2013/REC-prov-dm-20130430/ÿ

Huynh, TD, Stalla-Bourdillon, S. & Moreau, L. (2019). Spiegazioni basate sulla provenienza per decisioni automatizzate: rapporto finale del progetto IAA. URL : https://kclpure.kcl.ac.uk/portal/en/publications/provenancebased-explanations-forautomated-decisions(5b1426ce-d253-49fa-8390-4bb3abe65f54).html ÿ

Risorse per esplorare i tipi di algoritmi

Generale

Hastie, T., Tibshirani, R., Friedman, J. e Franklin, J. (2005). Gli elementi dell'apprendimento statistico: data mining, inferenza e previsione. *The Mathematical Intelligencer, 27(2),* 83-85. http://thuvien.thanglong.edu.vn:8081/dspace/bitstream/DHTL_123456789/4053/1/%5BSspringer%20Series%20in%20Statistics-1.pdf ÿ _

Molnar, C. (2019). Apprendimento automatico interpretabile: una guida per rendere spiegabili i modelli di scatole nere. https://christophm.github.io/interpretable-ml-book/ ÿ

Rudin, C. (2019). Smetti di spiegare i modelli di apprendimento automatico della scatola nera per decisioni ad alto rischio e usa invece modelli interpretabili. *Nature Machine Intelligence, 1(5)* , 206. https://www.nature.com/articles/s42256-019-0048-x ÿ

Regressione regolarizzata (LASSO e Ridge)

Gaines, BR e Zhou, H. (2016). Algoritmi per l'adattamento del lazo vincolato. *Journal of Computational and Graphical Statistics,* 27(4), 861-871. https://arxiv.org/pdf ÿ/1611.01511.pdf ÿ

Tibshirani, R. (1996). restringimento di regressione e la selezione tramite il lazo. Giornale della *Royal Statistical Society: Serie B (Metodologico),* 58(1), 267-288. http://beehive.cs.princeton.edu/course/read/tibshirani-jrssb-1996.pdf ÿ _

Modello lineare generalizzato (GLM)

https://CRAN.R-project.org/package=glmnetÿ _

Friedman, J., Hastie, T. e Tibshirani, R. (2010). *Percorsi di regolarizzazione per modelli lineari generalizzati tramite discesa di coordinate.* Giornale di software statistico, 33(1), 1-22. http://www.jstatsoft.org/v33/i01/ÿ _

Simon, N., Friedman, J., Hastie, T. e Tibshirani, R. (2011). *Percorsi di regolarizzazione per il modello dei rischi proporzionali di Cox tramite la discesa delle coordinate.* Giornale di software statistico, 39(5), 1-13. URL http://www.jstatsoft.org/v39/i05/ÿ _

Modello additivo generalizzato (GAM)

https://CRAN.R-project.org/package=gam

Lou, Y., Caruana, R. e Gehrke, J. (2012). Modelli intelligibili per la classificazione e la regressione. In *Atti della 18a conferenza internazionale ACM SIGKDD sulla scoperta della conoscenza e il data mining* (pp. 150-158). ACM. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.8241&rep=rep1&type=pdf ÿ

Legno, SN (2006). *Modelli additivi generalizzati: un'introduzione con* R. CRC Press.

Albero decisionale (DT)

Breiman, L., Friedman, J., Stone, CJ e Olshen, RA (1984). *Classificazione e alberi di regressione.* CRCPress.

Elenchi e insiemi di regole/decisioni

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M. e Rudin, C. (2017). Apprendimento di elenchi di regole certificabili per dati categoriali. *The Journal of Machine Learning Research,* 18(1), 8753-8830. http://www.jmlr.org/papers/volume18/17-716/17-716.pdf ÿ

Lakkaraju, H., Bach, SH e Leskovec, J. (2016, agosto). Set di decisioni interpretabili: un quadro comune per la descrizione e la previsione. In *Atti della 22a conferenza internazionale ACM SIGKDD sulla scoperta della conoscenza e il data mining* (pp. 1675-1684). ACM. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5108651/ÿ

Letham, B., Rudin, C., McCormick, TH e Madigan, D. (2015). Classificatori interpretabili mediante regole e analisi bayesiana: creazione di un modello di previsione dell'ictus migliore. *Gli annali della statistica applicata,* 9(3), 1350-1371. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1446488742ÿ

Wang, F. e Rudin, C. (2015). Liste di regole in calo. In *Intelligenza Artificiale e Statistica* (pp. 1013-1022). http://proceedings.mlr.press/v38/wang15a.pdf ÿ

Ragionamento basato su casi (CBR)/ Prototipo e critica

Aamodt, A. (1991). Un approccio integrato ad alta intensità di conoscenza alla risoluzione dei problemi e all'apprendimento sostenuto. *Gruppo di ingegneria della conoscenza ed elaborazione delle immagini. Università di Trondheim,* 27-85. http://www.dphu.org/uploads/attachments/books/books_4200_0.pdf ÿ

Aamodt, A., & Plaza, E. (1994). Ragionamento basato sui casi: problemi di base, variazioni metodologiche e approcci di sistema. *Comunicazioni AI,* 7(1), 39-59. https://www.idi.ntnu.no/emner/tdt4171/papers/AamodtPlaza94.pdf ÿ

Bichindaritz, I., & Marling, C. (2006). Ragionamento basato sui casi nelle scienze della salute: quali sono le prospettive?. *Intelligenza artificiale in medicina,* 36(2), 127-135. http://cs.oswego.edu/~bichinda/isc471-hci571/AIM2006.pdf ÿ

Bien, J. e Tibshirani, R. (2011). Selezione del prototipo per la classificazione interpretabile. *Gli Annali di Statistica Applicata,* 5(4), 2403-2424 ÿ.

Kim, B., Khanna, R. e Koyejo, OO (2016). Gli esempi non bastano, impara a criticare! critica per l'interpretabilità. In *Advances in Neural Information Processing Systems* (pp. 2280-2288).

http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf ÿ

MMD-critico in Python: https://github.com/BeenKim/MMD-critic ÿ

Kim, B., Rudin, C. e Shah, JA (2014). Il modello bayesiano del caso: un approccio generativo per il ragionamento basato sui casi e la classificazione del prototipo. In *Advances in Neural Information Processing Systems* (pp. 1952-1960). http://papers.nips.cc/paper/5313-the-bayesian-case-model-a-generative-approach-for-case-based-reasoning-and-prototype-classification.pdf ÿ

Modello intero lineare supersparso (SLIM)

Jung, J., Concannon, C., Shroff, R., Goel, S. e Goldstein, DG (2017). Regole semplici per decisioni complesse. *Disponibile presso SSRN 2919024.* https://arxiv.org/pdf/1702.04690.pdf ÿ

Rudin, C. e Ustun, B. (2018). Sistemi di punteggio ottimizzati: verso la fiducia nell'apprendimento automatico per l'assistenza sanitaria e la giustizia penale. *Interfacce,* 48(5), 449-466. https://pdfs.semanticscholar.org/b3d8/8871ae5432c84b76bf53f7316cf5f95a3938.pdf ÿ _

Ustun, B., & Rudin, C. (2016). Modelli interi lineari supersparsi per sistemi di punteggio medico ottimizzati. *Apprendimento automatico,* 102(3), 349-391. https://link.springer.com/article ÿ/10.1007/s10994-015-5528-6 ÿ

Sistemi di punteggio ottimizzati per problemi di classificazione in Python: https://github.com/ustunb/slim-python ÿ

Punteggi di rischio personalizzabili semplici in Python: https://github.com/ustunb/risk-slim ÿ

Risorse per esplorare strategie esplicative supplementari

Modelli surrogati (SM)

Bastani, O., Kim, C., & Bastani, H. (2017). Interpretabilità attraverso l'estrazione del modello. *arXiv prestampa arXiv:1706.09773.* https://obastani.github.io/docs/fatml17.pdf ÿ

Craven, M., & Shavlik, JW (1996). Estrazione di rappresentazioni strutturate ad albero di reti addestrate. In *Advances in neural information processing systems* (pp. 24-30). http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf ÿ _

Van Assche, A., & Blockeel, H. (2007). Vedere la foresta attraverso gli alberi: imparare un modello comprensibile da un insieme. Nella *conferenza europea sull'apprendimento automatico* (pp. 418-429). Springer, Berlino, Heidelberg. https://link.springer.com/content/pdf/10.1007/978-3-540-74958-5_39.pdf

Valdes, G., Luna, JM, Eaton, E., Simone II, CB, Ungar, LH e Solberg, TD (2016). MediBoost: uno strumento di stratificazione del paziente per un processo decisionale interpretabile nell'era della medicina di precisione. *Relazioni scientifiche,* 6, 37854. https://www.nature.com/articles/srep37854 ÿ

Grafico delle dipendenze parziali (PDP)

Friedman, JH (2001). Approssimazione della funzione Greedy: una macchina per aumentare il gradiente. *Annali di statistica,* 1189-1232. https://projecteuclid.org/download/pdf_1/euclid.aos/1013203451ÿ _

Greenwell, BM (2017). pdp: un pacchetto R per la costruzione di grafici a dipendenza parziale. *Il Giornale R,* 9(1), 421-436. https://pdfs.semanticscholar.org/cdfb/164f55e74d7b116ac63fc6c1c9e9cfd01cd8.pdf ÿ

Per il software in R: https://cran.r-project.org/web/packages/pdp/index.html ÿ

Grafico delle aspettative condizionali individuali (ICE)

Goldstein, A., Kapelner, A., Bleich, J. e Pitkin, E. (2015). Sbirciando all'interno della scatola nera: visualizzare l'apprendimento statistico con grafici delle aspettative condizionali individuali. *Journal of Computational and Graphical Statistics,* 24(1), 44-65. https://arxiv.org/pdf/1309.6392.pdf ÿ

Per il software in R vedere:
https://cran.r-project.org/web/packages/ICEbox/index.htmlÿ _ https://cran.r-project.org/web/packages/ICEbox/ICEbox.pdf ÿ

Complotti degli effetti locali accumulati (ALE)

Apley, DW e Zhu, J. (2019). Visualizzazione degli effetti delle variabili predittive nei modelli di apprendimento supervisionato con scatola nera. *arXiv prestampa arXiv:1612.08468.* https://arxiv.org/pdf ÿ/1612.08468; Visualizzazione ÿ

https://cran.r-project.org/web/packages/ALEPlot/index.html ÿ

Importanza variabile globale

Breiman, L. (2001). Foreste casuali. *Apprendimento automatico,* 45(1), 5-32 ÿ.

Casalicchio, G., Molnar, C., & Bischl, B. (2018, settembre). Visualizzazione dell'importanza delle caratteristiche per i modelli a scatola nera. Nella *conferenza europea congiunta sull'apprendimento automatico e la scoperta della conoscenza nei database* (pp. 655-670). Springer, Cham. https://arxiv.org/pdf/1804.06620.pdf ÿ

Fisher, A., Rudin, C. e Dominici, F. (2018). Tutti i modelli sono sbagliati, ma molti sono utili: apprendere l'importanza di una variabile studiando simultaneamente un'intera classe di modelli di previsione. arXiv:1801.01489 ÿ

Fisher, A., Rudin, C. e Dominici, F. (2018). Affidamento alla classe del modello: misure di importanza variabile per qualsiasi classe del modello di machine learning, dal punto di vista "Rashomon". *arXiv prestampa arXiv:1801.01489.* https://arxiv.org/abs/1801.01489v2 ÿ

Hooker, G. e Mentch, L. (2019). Si prega di interrompere la permuta delle funzioni: una spiegazione e alternative. *arXiv prestampa arXiv:1905.03151.* https://arxiv.org/pdf/1905.03151.pdf ÿ

Zhou, Z. e Hooker, G. (2019). Misurazione imparziale dell'importanza delle caratteristiche nei metodi ad albero. *arXiv prestampa arXiv:1903.05179.* https://arxiv.org/pdf/1903.05179.pdf ÿ

Interazione variabile globale

Friedman, JH e Popescu, BE (2008). Apprendimento predittivo attraverso set di regole. *Gli annali della statistica applicata,* 2(3), 916-954. https://projecteuclid.org/download/pdfview_1 ÿ/euclid.aoas/1223908046 ÿ

Greenwell, BM, Boehmke, BC e McCarthy, AJ (2018). Una misura di importanza variabile basata su modello semplice ed efficace. *arXiv prestampa arXiv:1805.04755.* https://arxiv.org/pdf/1805.04755.pdf ÿ

Hooker, G. (2004, agosto). Scoprire la struttura additiva nelle funzioni della scatola nera. In *Atti della decima conferenza internazionale ACM SIGKDD sulla scoperta della conoscenza e il data mining* (pp. 575-580). ACM. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7500&rep=rep1&type=pdf ÿ

Local Interpretable Model-Agnostic Explanation (LIME)

Ribeiro, MT, Singh, S. e Guestrin, C. (2016). Perché dovrei fidarmi di te?: Spiegare le previsioni di qualsiasi classificatore. In *Atti della 22a conferenza internazionale ACM SIGKDD sulla scoperta della conoscenza e il data mining* (pp. 1135-1144). ACM. https://arxiv.org/pdf/1602.04938.pdf?mod=article_inlineÿ _

LIME in Python: https://github.com/marcotcr/lime ÿ
Esperimenti LIME in Python: https://github.com/marcotcr/lime-experiments ÿ

Ribeiro, MT, Singh, S. e Guestrin, C. (2018). Ancoraggi: spiegazioni agnostiche del modello ad alta precisione. Nella *trentaduesima conferenza AAAI sull'intelligenza artificiale.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7500&rep=rep1&type=pdf ÿ _

Ancore in Python: https://github.com/marcotcr/anchor ÿ Ancora
esperimenti in Python: https://github.com/marcotcr/anchor-experiments ÿ

Shapley Additive ExPlanations (SHAP)

Lundberg, SM e Lee, SI (2017). Un approccio unificato all'interpretazione delle previsioni del modello. In *Advances in Neural Information Processing Systems* (pp. 4765-4774). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf ÿ

Software per SHAP e le sue estensioni in Python: https://github.com/slundberg/shap ÿ Wrapper
R per SHAP: https://modeloriented.github.io/shapper/ ÿ

Shapley, LS (1953). Un valore per i giochi per n persone. *Contributi alla teoria dei giochi,* 2(28), 307-317. http://www.library.fa.ru/files/Roth2.pdf#page=39 ÿ

Spiegazione controfattuale

Wachter, S., Mittelstadt, B. e Russell, C. (2017). Spiegazioni controfattuali senza aprire la scatola nera: decisioni automatizzate e GDPR. Harvey. JL & Tech., 31, 841. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf ÿ

Ustun, B., Spangher, A. e Liu, Y. (2019). Ricorso attuabile nella classificazione lineare. In *Atti della Conferenza sull'equità, la responsabilità e la trasparenza (pp.* 10-19). ACM. https://arxiv.org/pdf/1809.06514.pdf ÿ _

Valuta il ricorso nei modelli di classificazione lineare in Python: https://github.com/ustunb/actionable-recourse ÿ

Spiegatori secondari e sistemi basati sull'attenzione

Li, O., Liu, H., Chen, C. e Rudin, C. (2018). Deep learning per il ragionamento basato su casi tramite prototipi: una rete neurale che ne spiega le previsioni. Nella *trentaduesima conferenza AAAI sull'intelligenza artificiale.* https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17082/16552ÿ _

Park, DH, Hendricks, LA, Akata, Z., Schiele, B., Darrell, T. e Rohrbach, M. (2016). Spiegazioni attente: giustificare le decisioni e indicare le prove. *arXiv prestampa arXiv:1612.04757.* https://arxiv.org/pdf/1612.04757ÿ

Altre risorse per spiegazioni supplementari

Explainability 360 di IBM: http://aix360.mybluemix.net ÿ

Biecek, B. e Burzykowski, T. (2019). *Modelli predittivi: esplorare, spiegare ed eseguire il debug, machine learning interpretabile incentrato sull'uomo.* Estratto da https://pbiecek.github.io/PM_VEE/ ÿ

Software di accompagnamento, Dalex, Spiegazioni descrittive sull'apprendimento automatico : https://github.com/ModelOriented/DALEX ÿ

Przemysÿaw Biecek, *Risorse interessanti relative a XAI:* https://github.com/pbiecek/xai_resources ÿ

Christoph Molnar, iml: apprendimento automatico interpretabile https://cran.r-project.org/web/packages/iml/index.html ÿ

Allegato 5: Casi di assurance basata su argomenti

Un caso di assicurazione è un insieme di affermazioni strutturate, argomentazioni e prove che danno fiducia che un sistema di IA possiederà le qualità o le proprietà particolari che devono essere assicurate. Prendi, ad esempio, un caso di garanzia di "sicurezza e prestazioni". Ciò comporterebbe fornire un argomento, supportato da prove, che un sistema possiede le proprietà che gli consentiranno di funzionare in modo sicuro, protetto, affidabile, ecc. date le sfide del suo contesto operativo.

Sebbene i casi di assurance basati su argomenti siano storicamente emersi in domini critici per la sicurezza (come "casi di sicurezza" per le tecnologie basate su software), la metodologia è ampiamente utilizzata. Questo perché è un modo ragionevole per strutturare e documentare un approccio anticipatorio, basato sugli obiettivi e procedurale alla governance dell'innovazione. Ciò è in contrasto con i metodi più vecchi, più reattivi e prescrittivi che ora sono sempre più sfidati dal carattere complesso e in rapida evoluzione delle tecnologie di intelligenza artificiale emergenti.

Questo vecchio approccio prescrittivo sottolineava l'applicazione di standard generali universali che specificavano come dovrebbero essere costruiti i sistemi e spesso trattavano la governance come un esercizio di casella di controllo retrospettivo. Tuttavia, l'assicurazione basata su argomenti prende una strada diversa. Inizia all'inizio di un progetto di intelligenza artificiale e svolge un ruolo attivo in tutte le fasi del ciclo di vita della progettazione e dell'utilizzo. Inizia con obiettivi normativi di alto livello derivati dall'impatto ancorato al contesto e dalla valutazione basata sul rischio per ciascuna specifica applicazione di intelligenza artificiale, quindi espone argomentazioni strutturate che dimostrano:

1. come tali requisiti normativi affrontano gli impatti ei rischi associati all'uso del sistema nel suo ambiente operativo specificato;
2. in che modo le attività intraprese durante il flusso di lavoro di progettazione e distribuzione garantiscono le proprietà del sistema necessarie per la realizzazione di questi obiettivi; e
3. come sono state predisposte adeguate misure di monitoraggio e rivalutazione per garantire l'efficacia dei controlli attuati.

Abbiamo incluso un elenco di letture di base e risorse per aiutarti nell'area della governance e degli standard relativi ai casi di assicurazione basati su argomenti. Ciò include standard consolidati per l'assicurazione del sistema e del software dell'International Standards Organisation, della International Electrotechnical Commission e dell'Institute of Electrical and Electronics Engineers (serie ISO/IEC/IEEE 15026), nonché del Metamodel del caso di assicurazione strutturata (SACM) dell'Object Management Group). Ciò include anche riferimenti per molte delle principali piattaforme di assurance come Goal Structuring Notation (GSN), Claims, Arguments and Evidence Notation (CAE), NOR-STA Argumentation e Dynamic Safety Cases (DSC).

Componenti principali dei casi di assurance basati su argomenti

Sebbene esula dallo scopo di questa guida trattare i dettagli dei diversi metodi di creazione di casi di assicurazione, può essere utile fornire una visione ampia di come le componenti principali di qualsiasi caso di assicurazione globale si incastrano tra loro.

Obiettivi normativi di primo livello

Questi sono gli obiettivi o gli obiettivi di alto livello del sistema che affrontano i rischi e i potenziali danni che possono essere causati dall'uso di quel sistema nel suo ambiente operativo definito e necessitano quindi di garanzie. Nel contesto della spiegazione basata sui processi, questi includono equità e mitigazione dei pregiudizi,

responsabilità, sicurezza e prestazioni ottimali e impatto benefico e non dannoso. A partire da ciascuno di questi obiettivi normativi, la costruzione di un caso di garanzia implica quindi l'identificazione delle proprietà e delle qualità che un dato sistema deve possedere per garantire che raggiunga l'obiettivo specificato alla luce dei rischi e delle sfide che deve affrontare nel suo contesto operativo.

Affermazioni

Queste sono le proprietà, le qualità, i tratti o gli attributi che devono essere assicurati per realizzare gli obiettivi normativi di livello superiore. Ad esempio, in un caso di garanzia di equità e mitigazione dei pregiudizi, la proprietà "le variabili target o le loro proxy misurabili non riflettono pregiudizi strutturali o discriminazioni sottostanti" è una delle numerose affermazioni di questo tipo. In quanto componente centrale dell'argomentazione strutturata, deve essere supportata sia da argomentazioni di supporto appropriate su come le attività rilevanti alla base del processo di progettazione e sviluppo del sistema abbiano assicurato che i pregiudizi strutturali non fossero, di fatto, incorporati nelle variabili obiettivo, sia da prove corrispondenti che hanno documentato queste attività.

In alcuni metodi di argomentazione strutturata come GSN, le affermazioni sono qualificate da **componenti di contesto,** che:

- chiarire la portata di una determinata affermazione;
- fornisce definizioni e informazioni di base;
- rendere esplicite le ipotesi rilevanti sul sistema e sull'ambiente; e
- esplicitare i rischi e le esigenze di mitigazione dei rischi associati al reclamo durante la progettazione e il ciclo di vita operativo del sistema.

Le affermazioni possono essere qualificate anche da **componenti di giustificazione,** ovvero chiarimenti di:

- perché le affermazioni sono state scelte; e
- come forniscono una soluzione o mezzi di realizzazione per l'obiettivo normativo specificato.

In generale, l'aggiunta di componenti di contesto e giustificazione rafforza l'accuratezza e la completezza delle affermazioni, consentendo loro di supportare gli obiettivi di primo livello del sistema. Questa attenzione alla precisione, al chiarimento e alla completezza è fondamentale per stabilire la fiducia attraverso lo sviluppo di un'assicurazione efficace
scatola.

argomenti

Questi supportano le affermazioni collegandole con prove e altre affermazioni di supporto attraverso il ragionamento.
Gli argomenti forniscono garanzie per le affermazioni stabilendo una relazione inferenziale che collega la proprietà proposta con un corpo di prove e un supporto argomentativo sufficienti a stabilirne l'accettabilità razionale o la verità. Ad esempio, in un caso di sicurezza e garanzia delle prestazioni, che aveva "il sistema è sufficientemente robusto" come una delle sue affermazioni, un possibile argomento potrebbe essere "i processi di formazione includevano un elemento di aumento in cui sono stati impiegati esempi contraddittori e disturbi per modellare il mondo reale duro termini'. Tale affermazione sarebbe quindi supportata dal supporto probatorio che ciò sia effettivamente accaduto durante la progettazione e lo sviluppo del modello di intelligenza artificiale.

Sebbene le argomentazioni giustificate siano sempre sostenute da un corpus di prove, possono anche essere supportate da affermazioni e **presupposti subordinati** (cioè affermazioni senza ulteriore supporto che sono considerate ovvie o vere). Le rivendicazioni subordinate sottoscrivono argomentazioni (e le affermazioni di livello superiore che supportano) basate sulle proprie argomentazioni e prove. Nell'argomento strutturato, ci saranno spesso più livelli subordinati

affermazioni, che lavorano insieme per fornire giustificazione per l'accettabilità razionale o la verità delle affermazioni di livello superiore.

Evidenza

Questa è la raccolta di manufatti e documentazione che forniscono supporto probatorio per le affermazioni avanzate nel caso di assicurazione. Un corpo di prove è formato da informazioni oggettive, dimostrabili e ripetibili registrate durante la produzione e l'uso di un sistema. Ciò è alla base delle argomentazioni che giustificano le affermazioni di assurance. In alcuni casi, un corpus di prove può essere organizzato in un **archivio di prove** (SACM) in cui è possibile accedere a tali informazioni primarie insieme a informazioni secondarie sulla gestione delle prove, sull'interpretazione delle prove e sul chiarimento del supporto probatorio alla base delle affermazioni dell'assicurazione
scatola.

Vantaggi dell'approccio alla spiegazione basata sul processo attraverso casi di assurance basati su argomenti

Ci sono diversi vantaggi nell'usare la garanzia basata su argomenti per organizzare la documentazione delle tue pratiche di innovazione per spiegazioni basate sui processi:

- I casi di assicurazione richiedono una comprensione proattiva e end-to-end degli impatti e dei rischi derivanti da ciascuna specifica applicazione di intelligenza artificiale. La loro effettiva esecuzione è ancorata alla creazione di controlli pratici che dimostrino che tali impatti e rischi sono stati gestiti in modo appropriato. Per questo motivo, i casi di assurance incoraggiano l'integrazione pianificata di buoni controlli di governance basati sui processi. Questo, a sua volta, garantisce che gli obiettivi che regolano lo sviluppo dei sistemi di IA siano stati raggiunti, con un metodo di garanzia documentata deliberato e basato su argomenti che lo dimostri. Nella garanzia basata su argomenti, i processi di governance e documentazione anticipati e orientati agli obiettivi lavorano di pari passo, rafforzando reciprocamente le migliori pratiche e migliorando la qualità dei prodotti e dei servizi che supportano.

- La garanzia basata su argomenti implica un metodo di pratica di governance basata sul ragionamento piuttosto che un insieme di istruzioni specifiche per attività o tecnologia. Ciò consente ai progettisti e agli sviluppatori di intelligenza artificiale di affrontare una vasta gamma di attività di governance con un unico metodo di utilizzo di argomenti strutturati per garantire proprietà che soddisfano i requisiti standard e mitigano i rischi. Ciò significa anche che le procedure per la creazione di casi di assicurazione sono uniformi e che la loro documentazione può essere più facilmente standardizzata e automatizzata (come si vede, ad esempio, in varie piattaforme di assicurazione come GSN, SACM, CAE e NOR-STA Argumentation).

- La garanzia basata su argomenti può consentire un'efficace comunicazione multi-stakeholder. Può generare fiducia nel fatto che una determinata applicazione di intelligenza artificiale possieda le proprietà desiderate sulla base di motivi espliciti, ben motivati e supportati da prove. Quando sono fatti in modo efficace, i casi di assurance trasmettono informazioni in modo chiaro e preciso a vari gruppi di stakeholder attraverso argomentazioni strutturate. Questi dimostrano che gli obiettivi specificati sono stati raggiunti e i rischi sono stati mitigati. Lo fanno fornendo prove documentali che le proprietà del sistema necessarie per raggiungere questi obiettivi sono state assicurate da solide argomentazioni.

- L'utilizzo della metodologia di assurance basata su argomenti può consentire di personalizzare i casi di assurance e adattarli al pubblico pertinente. Questi casi di assurance sono costruiti su argomenti strutturati (affermazioni, giustificazioni e prove) in linguaggio naturale, quindi sono più facilmente comprensibili da coloro che non sono specialisti tecnici. Mentre argomentazioni tecniche dettagliate e prove possono supportare casi di assurance, la base di questi casi nel ragionamento quotidiano li rende particolarmente suscettibili di sintesi non tecniche e comprensibili. Una sintesi di un caso di garanzia fornito a un destinatario di una decisione può quindi essere supportata da una versione più dettagliata che include argomentazioni strutturali estese meglio adattate a esperti, valutatori indipendenti e revisori dei conti. Allo stesso modo, le prove utilizzate per un caso di assurance possono esserlo

organizzato per adattarsi al pubblico e al contesto della spiegazione. In questo modo, tutte le informazioni potenzialmente commercialmente sensibili o lesive della privacy, che costituiscono una parte del corpus di prove, possono essere conservate in un archivio di prove. Questo può essere reso accessibile a un pubblico più ristretto di supervisori, valutatori e revisori interni o esterni.

Governare le pratiche di approvvigionamento e gestire le aspettative degli stakeholder attraverso soluzioni personalizzate assicurazione

Sia per i fornitori che per i clienti (cioè sviluppatori e utenti/fornitori), l'assicurazione basata su argomenti può fornire un modo ragionevole per governare le pratiche di approvvigionamento e per gestire reciprocamente le aspettative. Se un fornitore ha un approccio deliberato e anticipatorio alla progettazione del sistema di IA consapevole delle spiegazioni richiesto dall'assicurazione basata su argomenti, sarà in grado di assicurare meglio (attraverso giustificazione, evidenza e documentazione) le proprietà cruciali dei suoi modelli a coloro che sono interessati ad acquisire loro.

Offrendo in anticipo un portafoglio di garanzie supportate da prove, un fornitore sarà in grado di dimostrare che i propri prodotti sono stati progettati tenendo conto degli obiettivi normativi appropriati. Saranno anche in grado di assicurare ai potenziali clienti che questi obiettivi sono stati raggiunti attraverso i processi di sviluppo. Ciò consentirà quindi anche agli utenti/appaltatori di trasmettere questa parte della spiegazione basata sul processo ai destinatari delle decisioni e alle parti interessate. Consentirebbe inoltre agli appaltatori di valutare in modo più efficace se il portafoglio di garanzie soddisfa i criteri normativi e gli standard di innovazione dell'IA che stanno cercando in base alla loro cultura organizzativa, contesto di dominio e interessi applicativi.

L'utilizzo di casi di assurance consentirà inoltre una valutazione indipendente basata su standard e audit di terze parti. I compiti di gestione e condivisione delle informazioni possono essere svolti in modo efficiente tra sviluppatori, utenti, valutatori e revisori dei conti. Ciò può fornire una piattaforma comune e consolidata per la spiegazione basata sui processi, che organizza sia la presentazione dei dettagli dei casi di assurance sia l'accessibilità delle informazioni che li supportano. Ciò semplificherà i processi di comunicazione tra tutte le parti interessate interessate, preservando al contempo gli aspetti che generano fiducia della trasparenza procedurale e organizzativa fino in fondo.

Ulteriori letture

Risorse per esplorare la documentazione e l'assicurazione basata su argomenti

Letture generali sulla documentazione per la progettazione e l'implementazione responsabile dell'IA

Schede informative ÿ

Schede tecniche per set di dati ÿ

Schede modello per la creazione di report sui modelli ÿ

Blog sulla struttura di auditing AI ÿ

Comprendere l'etica e la sicurezza dell'intelligenza artificiale ÿ

Norme e regolamenti pertinenti sull'assicurazione basata su argomenti e sui casi di sicurezza

ISO/IEC/IEEE 15026-1:2019, *Ingegneria dei sistemi e del software — Assicurazione dei sistemi e del software — Parte 1: Concetti e vocabolario.*

ISO/IEC 15026-2:2011, *Ingegneria dei sistemi e del software — Assicurazione dei sistemi e del software — Parte 2: Caso di garanzia.*

ISO/IEC 15026-3: 2015, *Ingegneria dei sistemi e del software — Garanzia dei sistemi e del software — Parte 3: Livelli di integrità del sistema.*

ISO/IEC 15026-4: 2012, *Ingegneria dei sistemi e del software — Assicurazione dei sistemi e del software — Parte 4: Assicurazione nel ciclo di vita.*

Object Management Group, *Structured Assurance Case Metamodel* (SACM), versione 2.1 beta, marzo 2019.

Ministero della Difesa. Standard di difesa 00-42 Edizione 2, Linee guida sulla garanzia di affidabilità e manutenibilità (R&M). Parte 3, Caso R&M, 6 giugno 2003.

Ministero della Difesa. Standard di difesa 00-55 (PARTE 1)/Edizione 4, *Requisiti per il software relativo alla sicurezza nelle apparecchiature per la difesa, parte 1: Requisiti,* dicembre 2004.

Ministero della Difesa. Standard di difesa 00-55 (PARTE 2)/Edizione 2, *Requisiti per il software relativo alla sicurezza nelle apparecchiature per la difesa, parte 2: Linee guida,* 21 agosto 1997.

Ministero della Difesa. Standard di difesa 00-56. *Requisiti di gestione della sicurezza per i sistemi di difesa. Parte 1. Requisiti Edizione* 4, 01 giugno 2007

Ministero della Difesa. Standard di difesa 00-56. *Requisiti di gestione della sicurezza per i sistemi di difesa. Parte 2: Guida per stabilire un mezzo per conformarsi alla Parte 1 Edizione* 4, 01 giugno 2007.

UK CAA CAP 760 *Guida alla condotta dell'identificazione dei pericoli, della valutazione del rischio e della produzione di casi di sicurezza per gli operatori aeroportuali e i fornitori di servizi di traffico aereo,* 13 gennaio 2006.

I regolamenti sulle installazioni offshore (caso di sicurezza) 2005 n. 3117. http://www.legislation.gov.uk/uksi/2005/3117/contents/made ÿ

Il controllo dei rischi di incidenti rilevanti (modifica) Regolamenti 2005 n. 1088. http://www.legislation.gov.uk/uksi/2005/1088/contents/made

Esecutivo per la salute e la sicurezza. *Principi di valutazione della sicurezza per gli impianti nucleari.* HSE; 2006.

Le ferrovie e altri sistemi di trasporto guidato (sicurezza) regolamenti 2006. UK Statutory Instrument 2006 No.599. http://www.legislation.gov.uk/uksi/2006/599/contents/madeÿ _

Direttiva CE 91/440/CEE. *Sullo sviluppo delle ferrovie comunitarie.* 29 luglio 1991.

Letture di base sui metodi di assurance basata su argomenti

Ankrum, TS e Kromholz, AH (2005). Casi di assicurazione strutturata: tre standard comuni. In

Nono IEEE International Symposium on High-Assurance Systems Engineering (HASE'05) (pp. 99-108).

Ashmore, R., Calinescu, R. e Paterson, C. (2019). Assicurare il ciclo di vita dell'apprendimento automatico: Desiderata, metodi e sfide. *arXiv prestampa arXiv:1905.04223.*

Barry, signor (2011). CertWare: un banco di lavoro per la produzione e l'analisi di casi di sicurezza. Nel *2011 conferenza aerospaziale* (pp. 1-10). IEEE.

Bloomfield, R. e Netkachova, K. (2014). Elementi costitutivi per i casi assicurativi. Nel *2014 IEEE International Symposium on Software Reliability Engineering Workshops (pp.* 186-191). IEEE.

Bloomfield, R. e Bishop, P. (2010). Casi di sicurezza e assicurazione: passato, presente e possibile futuro: una prospettiva di Adelard. In *Rendere i sistemi più sicuri* (pp. 51-67). Springer, Londra.

Cârlan, C., Barner, S., Diewald, A., Tsalidis, A. e Voss, S. (2017). ExplicitCase: sviluppo integrato basato su modelli di sistemi e casi di sicurezza. *Conferenza internazionale sulla sicurezza, l'affidabilità e la sicurezza dei computer* (pp. 52-63). Springer.

Denney, E. e Pai, G. (2013). Una base formale per i modelli di casi di sicurezza. Nella *conferenza internazionale sulla sicurezza, affidabilità e sicurezza dei computer* (pagine 21-32). Springer.

Denney, E., Pai, G. e Habli, I. (2015). Casi di sicurezza dinamica per l'assicurazione della sicurezza per tutta la vita. *IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 2, pp. 587-590). IEEE.

Denney, E. e Pai, G. (2018). Strumento di supporto per lo sviluppo di casi di assurance. *Ingegneria del software automatizzata,* 25(3), 435-499.

Despoto, G. (2004). Estendere il concetto di custodia di sicurezza per affrontare la dipendenza. *Atti della 22a Conferenza Internazionale sulla Sicurezza dei Sistemi.*

Gacek, A., Backes, J., Cofer, D., Slind, K. e Whalen, M. (2014). Risoluto: un linguaggio di casi di garanzia per i modelli di architettura. *ACM SIGAda Ada Lettere,* 34(3), 19-28.

Ge, X., Rijo, R., Paige, RF, Kelly, TP e McDermid, JA (2012). Introduzione alla notazione di strutturazione degli obiettivi per spiegare le decisioni nella pratica clinica. *Tecnologia Procedia,* 5, 686-695.

Gleirscher, M. e Kugele, S. (2019). Garanzia della sicurezza del sistema: un'indagine sui modelli di progettazione e argomenti. *arXiv prestampa arXiv:1902.05537.*

Górski, J., Jarzÿbowicz, A., Miler, J., Witkowicz, M., Czyÿnikiewicz, J. e Jar, P. (2012). Supportare l'assicurazione mediante servizi di argomentazione basata sull'evidenza. *Conferenza internazionale sulla sicurezza, l'affidabilità e la sicurezza dei computer* (pp. 417-426). Springer, Berlino, Heidelberg.

Habli, I. e Kelly, T. (2014, luglio). Bilanciare gli argomenti formali e informali nei casi di sicurezza. In *VeriSure: Verification and Assurance Workshop, in collaborazione con Computer-Aided Verification (CAV).*

Hawkins, R., Habli, I., Kolovos, D., Paige, R. e Kelly, T. (2015). Intrecciare un caso di garanzia dalla progettazione: un approccio basato su modelli. *2015 IEEE 16th International Symposium on High Assurance Systems Engineering* (pp. 110-117). IEEE.

Fondazione per la salute (2012). Evidenza: utilizzo di casi di sicurezza nell'industria e nell'assistenza sanitaria.

Kelly, T. (1998) *Argomentare la sicurezza: un approccio sistematico alla gestione dei casi di sicurezza. Tesi di dottorato.* Università di York: Dipartimento di Informatica.

Kelly, T. e McDermid, J. (1998). Modelli di casi di sicurezza: riutilizzo di argomenti di successo. *Colloquio IEEE sulla comprensione dei modelli e la loro applicazione all'ingegneria dei sistemi, Londra.*

Kelly, T. (2003). *Un approccio sistematico alla gestione dei casi di sicurezza.* SAE Internazionale.

Kelly, T. e Weaver, R. (2004). La notazione di strutturazione dell'obiettivo: una notazione di argomento di sicurezza. In *Atti dei sistemi e delle reti dipendenti 2004 workshop sui casi di assicurazione.*

Maksimov, M., Fung, NL, Kokaly, S. e Chechik, M. (2018). Due decenni di strumenti per i casi assicurativi: un'indagine. *Conferenza internazionale sulla sicurezza, l'affidabilità e la sicurezza dei computer* (pp. 49-59). Springer.

Nemouchi, Y., Foster, S., Gleirscher, M. e Kelly, T. (2019). Casi di assicurazione meccanizzata con metodi formali integrati in Isabelle. *arXiv prestampa arXiv:1905.06192.*

Netkachova, K., Netkachov, O. e Bloomfield, R. (2014). Strumento di supporto per gli elementi costitutivi del caso assicurativo. Nella *conferenza internazionale sulla sicurezza, l'affidabilità e la sicurezza dei computer* (pp. 62-71). Springer.

Picardi, C., Hawkins, R., Paterson, C. e Habli, I. (2019, settembre). Un modello per argomentare la garanzia dell'apprendimento automatico nei sistemi di diagnosi medica. Nella *conferenza internazionale sulla sicurezza, l'affidabilità e la sicurezza dei computer* (pp. 165-179). Springer.

Picardi, C., Paterson, C., Hawkins, RD, Calinescu, R. e Habli, I. (2020). Modelli e processi di argomenti di garanzia per l'apprendimento automatico nei sistemi relativi alla sicurezza. *Atti del Workshop sulla Sicurezza dell'Intelligenza Artificiale* (pp. 23-30). Atti del seminario CEUR.

Rushby, J. (2015). L'interpretazione e la valutazione dei casi assicurativi. comp. Laboratorio Scientifico, SRI International, Tech. Rappresentante. SRI-CSL-15-01.

Strunk, EA e Knight, JC (2008). La sintesi essenziale di strutture problematiche e casi di assurance. *Sistemi esperti,* 25(1), 9-27.