UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

ExOn: una nuova ontologia sui metodi per l'Explainable Artificial Intelligence

ExOn: a new ontology on Explainable Artificial Intelligence methods

Relatore:
Prof. Primo Zingaretti

Candidato:
Alessandro Muscatello

Anno accademico 2022/2023

Indice

Sommario **II**

1 Introduzione **1**

2 Motivazioni **3**

3 Stato dell'arte **6**
3.1 Storia dei metodi di XAI: le tre generazioni 6
3.2 Le classificazioni dei metodi 8
3.3 Terminologia 8

4 ExOn: una nuova ontologia **12**
4.1 Le ontologie 12
4.2 Le classi di ExOn 14
4.3 Le proprietà dell'ontologia 14
4.3.1 Le object property 20
4.3.2 Le data property 22
4.4 Strumenti utilizzati 23
4.5 Esempi di ricerca 24

5 Discussione **26**
5.1 Confronto con Arrieta et al. 26
5.2 Confronto con Speith 27
5.3 Confronto con Nauta et al. 27

6 Conclusioni e sviluppi futuri **30**
6.1 Sviluppi futuri 31

7 Riferimenti bibliografici **32**

Sommario

Il tema dell'Intelligenza Artificiale (IA) è senza dubbio uno dei più discussi negli ultimi anni. L'impiego dell'IA si estende in diversi settori, con una crescente popolarità soprattutto in ambiti critici come la sanità e la sicurezza. Da cui nasce la necessità di garantire trasparenza e comprensibilità nei processi decisionali automatizzati al fine di migliorare l'affidabilità, la responsabilità e la fiducia nelle applicazioni IA. In proposito, l'Unione Europea è in procinto di approvare la nuova normativa AiAct, che esplicita tali preoccupazioni sull'utilizzo di IA affidabili e trasparenti. I metodi legati all'Explainable Artificial Intelligence (XAI) cercano di spiegare le ragioni alla base delle risposte fornite da un sistema automatico. Tuttavia, è evidente dai lavori presenti nello stato dell'arte che non esiste una classificazione universalmente accettata che copra tutti gli aspetti dell'argomento. Questa lacuna è dovuta non solo alla grande quantità di nuovi metodi proposti, ma anche alla diversità di funzionamento tra di essi.

In generale esistono delle tassonomie, ovvero strutture che categorizzano e organizzano le diverse tecniche, approcci e concetti relativi all'explainability delle intelligenze artificiali. Le tassonomie possono essere limitate dalla loro struttura rigida e dalla difficoltà di rappresentare relazioni complesse tra i concetti. In alcuni casi, può avvenire che alcune metodologie non siano adattabili facilmente ad una struttura gerarchica.

Un'ontologia, invece, è una forma più avanzata di rappresentazione concettuale rispetto a una tassonomia, dato che va oltre la semplice categorizzazione e offre una descrizione dettagliata delle relazioni tra i concetti. In altre parole, un'ontologia non solo classifica i concetti, ma stabilisce anche le connessioni semantiche tra di essi. Per questo motivo viene proposta ExOn, una nuova ontologia, in grado di creare una struttura formale per categorizzare le pubblicazioni scientifiche sull'XAI e di incorporare le informazioni essenziali in modo accessibile sia per gli utenti che per i ricercatori neofiti in questo campo.

Rispetto alle classificazioni esistenti in letteratura, ExOn rappresenta un avanzamento, trasformando il concetto di classificazione da una semplice tassonomia dei metodi a una formalizzazione sotto forma di ontologia. Inoltre, le categorie utilizzate sono state derivate dalle più recenti categorizzazioni al fine di coprire ogni aspetto dei metodi. ExOn facilita l'identificazione dei collegamenti e delle influenze tra le pubblicazioni, grazie alla formalizzazione dei concetti di 'citazione' e 'metodo derivato da', migliorando la tracciabilità della ricerca.

Infine, considerando la capacità intrinseca di espansione delle ontologie, viene discusso

il possibile ampliamento per formalizzare il concetto di autore delle pubblicazioni e l'integrazione con l'ontologia ANNETT-O per migliorare ulteriormente la rappresentazione delle reti neurali profonde. Inoltre, si suggerisce la valutazione di un'applicazione con interfaccia grafica per semplificare la navigazione, mantenendo al contempo la formalità dell'ontologia.

Di seguito verranno elencati i capitoli della tesi con una breve descrizione del contenuto:

1. **Introduzione**: introduzione nel mondo dell'XAI e al lavoro proposto.
2. **Motivazioni**: le motivazioni della creazione di ExOn.
3. **Stato dell'arte**: storia dell'XAI, terminologia e classificazioni esistenti.
4. **ExOn: una nuova ontologia**: presentazione di ExOn, la sua struttura, le proprietà, ed esempi di navigazione.
5. **Discussione**: confronto tra le classificazioni esistenti ed ExOn.
6. **Conclusione e sviluppi futuri**: riepilogo dei risultati ottenuti e sviluppi futuri.

1 Introduzione

L'argomento dell'IA (Intelligenza Artificiale) è sicuramente uno dei maggiori argomenti di discussione degli ultimi anni. La sua importanza è tale da aver spinto l'Unione Europea a stanziare 180 milioni di Euro per promuovere la ricerca nell'ambito dell'intelligenza artificiale e robotica intelligente [10], investimento facente parte del progetto "Digital Europe" con un budget complessivo di 7,5 miliardi di Euro [24]. Più recentemente il progetto di finaziamenti "Horizon Europe" ha messo a disposizione un totale di 95,5 miliardi di Euro, il più grande investimento sulla ricerca mai introdotto [8].

L'utilizzo della IA non si limita solo alla robotica, ma si estende in molti altri campi come: riconoscimento dei tentativi di accesso illecito attraverso spoofing di sistemi di riconoscimento facciale [17], auto a guida autonoma, lotta agli attacchi informatici, utilizzo come supporto alle decisioni in campo medico, assistenti digitali o traduzione automatica [7]. Molti strumenti sono stati pubblicati e resi disponibili agli utenti pur non necessitando di avere le conoscenze tecniche sul funzionamento di tali sistemi. Dunque rimane compito dei programmatori avere cura di creare strumenti che siano affidabili, di tenere in conto la pericolosità dell'ambiente in cui verranno utilizzati, e fornire garanzie che il prodotto fornito funzioni correttamente. Tali strumenti automatici stanno diventando sempre più popolari in ambiti critici come la sanità o la sicurezza [12], situazioni in cui le azioni intraprese sono giudicate sotto il punto di vista etico e morale; in pratica ogni azione deve essere giustificata. Oltre agli investimenti che ha deciso di promuovere, l'UE ha anche fornito delle indicazioni sulla necessità di creare e utilizzare solo intelligenze artificiali che garantiscano l'affidabilità, la sicurezza, e l'eticità delle loro risposte; oltre alla suddivisione in quattro categorie di rischio in base agli specifici ambiti di utilizzo [4].

Per affrontare tale problematica è necessario che i sistemi automatici che si basano sull'IA abbiano in qualche modo la possibilità di spiegare le motivazioni delle loro risposte, ad esempio nel caso di un rifiuto di una richiesta per un prestito bancario si devono poter giustificare le motivazioni del rifiuto e valutare se esse non siano state influenzate da parametri eticamente scorretti, come il colore della pelle. I metodi di XAI (eXplainable Artificial Intelligence) tentano di fornire una spiegazione sulle motivazioni che hanno portato un sistema a fornire tali risposte. Questo ambito nasce dal momento in cui i sistemi di IA sono basati su tecniche di ML (Machine Leaning) che per loro natura sono opache, o di DL (Deep Learning) che sono ancora più complesse, tanto da essere chiamate tecniche "black-box".

In questa dissertazione verrà mostrato il problema della classificazione dei metodi di XAI attraverso le classificazioni presenti nello stato dell'arte, poi verranno elencate una serie di definizioni sulla terminologia corrente dell'ambito evidenziando il problema il problema dell'assenza ne di una classificazione comune ne di una terminologia comune. Quindi viene presentata una nuova ontologia, ExOn, che pone come obiettivo la creazione di una struttura formale in grado di poter definire ed accogliere in maniera esaustiva, attraverso la sua generalità, le pubblicazioni scientifiche sull'argomento dell'XAI, e tale da permettere l'inserimento delle informazioni essenziali delle pubblicazioni; ma allo stesso tempo sufficientemente semplice da essere navigata da utenti e ricercatori che hanno da poco o non hanno mai esplorato questo ambito.

Questo nuovo strumento vuole fornire un supporto a chi, ad esempio, ha prodotto un modello di intelligenza artificiale e vuole migliorare la fiducia del proprio modello applicando una tecnica di XAI; oppure a coloro che sono interessati a creare un nuovo modello di XAI e quindi vogliono avere uno strumento che permetta loro di avere non solo una panoramica generale sull'argomento ma che possa anche fornire dettagli sulle principali correnti, e un'indicazione su quali pubblicazioni hanno più influenzato un certo metodo.

Con la presentazione di ExOn, dunque, si vuole creare uno strumento in grado di:

- Fornire supporto a ricercatori ed utenti interessati all'argomento dell'XAI.
- Fornire una visuale personalizzata in base alla categoria di interesse dell'utilizzatore.
- Evidenziare i collegamenti e le influenze che ci sono state tra le pubblicazioni, identificando le correnti più emergenti e promettenti.
- Migliorare lo stato dell'arte delle classificazioni utilizzando la concettualizzazione più recente dell'ambito e sfruttando una definizione formale.

2 Motivazioni

L'interesse della comunità scientifica riguardo gli argomenti dell'Explainable Artificial Intelligence hanno visto una notevole crescita negli ultimi anni, ad esempio si veda l'aumento delle pubblicazioni su ArXiv (Figura 1) che abbiano nel titolo o nell'abstract almeno una keyword[1] compatibile con l'ambito dell'XAI. Questo è dovuto all'incredibile impatto che le soluzioni proposte da alcune aziende, come Watson di IBM o ChatGPT di OpenAI, abbiano notevolmente fatto accendere i riflettori in questo settore. E ciò ha stimolato la preoccupazione della possibilità di un utilizzo non appropriato di alcuni sistemi che possono essere considerati come pericolosi. La normativa europea [4], già citata precedentemente, pur non essendo ancora approvata mette in risalto alcuni punti fondamentali sulle caratteristiche che un sistema automatico che sfrutti una IA deve avere. Di seguito si riporta un breve elenco di preoccupazioni che sono state espresse dal legislatore riguardo ai possibili pericoli che un sistema di IA può portare se non gestito correttamente:

- Si necessita l'utilizzo di dati di alta qualità per l'addestramento con una attenta governance dei dati, in particolare quando vengono trattati dati personali o di rilevante interesse pubblico.
- I sistemi non devono permettere pratiche di manipolazione, sfruttamento e controllo sociale.
- I sistemi non devono distorcere materialmente il comportamento di una persona ai fini di provocare danno a tale persona o altre.
- I sistemi non devono violare il diritto alla dignità e alla non discriminazione.
- I sistemi potrebbero avere una ripercussione negativa per la salute e la sicurezza, per cui bisogna fare particolare attenzione in situazioni in cui tali sistemi siano integrati come una componente di un prodotto.

[1]Keywords: explainable ai, explainable machine learning, explainable aiml, transparent ai, transparent machine learning, transparent aiml, interpretable ai, interpretable machine learning, interpretable aiml, xai, black-box ai, black-box machine learning, black-box aiml, black box ai, black box machine learning, black box aiml, opaque ai, opaque machine learning, opaque aiml, uninterpretable ai, uninterpretable machine learning, uninterpretable aiml, non-transparent ai, non-transparent machine learning, non-transparent aiml, inexplicable ai, inexplicable machine learning, inexplicable aiml, unexplainable ai, unexplainable machine learning, unexplainable aiml

Nella normativa vengono inoltre definiti una serie di situazioni in cui l'utilizzo di sistemi di IA devono essere considerate ad "alto rischio". In tali situazioni deve venir posta maggiore attenzione da parte degli sviluppatori e dei fornitori a causa dell'elevata possibilità di poter arrecare danni a persone o cose. Di seguito viene riportato un breve elenco delle situazioni che sono state valutate ad alto rischio e per le quali si necessita una maggiore attenzione:

- Identificazione biometrica in 'tempo reale'.
- Gestione della sicurezza delle infrastrutture critiche come forniture di acqua, gas, elettricità o traffico stradale.
- Valutazione della prestazione dei partecipanti a corsi di formazione ed istruzione.
- Selezione del personale in candidature lavorative.
- Valutazione del diritto a concedere, ridurre, revocare o recuperare prestazioni o servizi di assistenza pubblici e privati.
- Valutazione creditizia bancaria.
- Valutazione individuali di persone per il rischio di reato o di recidiva.
- Valutazione dello stato emotivo di persone (poligrafo).
- Valutazione di priorità per l'invio di servizi di soccorso di emergenza.
- Assistenza alle autorità giudiziarie nell'interpretazione dei fatti e nell'applicazione delle normative.

Oltre ad affrontare queste problematiche di fondamentale importanza, vi è comunque presente la preoccupazione sulla trasparenza dei sistemi di IA. La natura opaca dei sistemi di Deep Learning rende difficile comprendere i funzionamenti interni dei modelli. Viene quindi riportato un passo fondamentale dell'AI Act riguardo la trasparenza:

> *(43) Per ovviare all'opacità che può rendere alcuni sistemi di IA incomprensibili o troppo complessi per le persone fisiche, è opportuno imporre un certo grado di trasparenza per i sistemi di IA ad alto rischio. Gli utenti dovrebbero poter interpretare gli output del sistema e utilizzarlo in modo adeguato. I sistemi di IA ad alto rischio dovrebbero pertanto essere corredati di documentazione e*

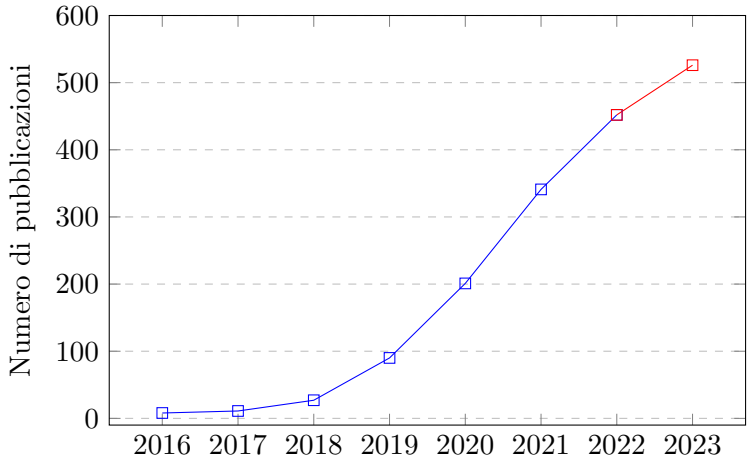

Figura 1: Numero di pubblicazioni su ArXiv dal 2016 a Luglio 2023 che hanno nel titolo o nell'abstract almeno una keyword per l'XAI [25]

> *istruzioni per l'uso pertinenti, nonché di informazioni concise e chiare, anche in relazione, se del caso, ai possibili rischi in termini di diritti fondamentali e discriminazione.*

Al fine di permettere il soddisfacimento non solo delle regolamentazioni che verranno emanate con l'AI Act, ma per fornire un prodotto di qualità ai propri consumatori, questa tesi propone un nuovo strumento, ExOn, che possa essere di supporto ai programmatori e alle aziende che vorranno immettere nel mercato dei sistemi di qualità, avendo pieno controllo sul funzionamento, spesso nascosto, delle IA. ExOn sarà in grado di fornire indicazioni nel panorama delle pubblicazioni per le tecniche di XAI, volte proprio ad migliorare la conscenza del funzionamento delle IA e aumentando la trasparenza dei sistemi.

3 Stato dell'arte

L'utilizzo di diversi metodi matematici e permettono ai sistemi di ML e DL di estrarre informazioni su un enorme quantitativo di dati con prestazioni straordinarie se paragonate alle capacità umane. Tale capacità di apprendere i pattern nei dati ha però come svantaggio che le regole apprese non sono direttamente visualizzabili e accessibili. Ciò comporta che non è possibile determinare con certezza se gli schemi che il sistema impara attraverso i dati siano effettivamente validi, o solo una coincidenza della particolarità del dataset utilizzato o della particolare casualità dell'addestramento. In tal senso la fase di 'testing' dei modelli ha l'obiettivo di valutare le performance e la capacità predittiva del modello. Questa fase è critica per garantire che il modello sia in grado di generalizzare bene su dati precedentemente non visti e che funzioni correttamente nell'applicazione finale. L'obiettivo della ricerca dei metodi di XAI è quello di migliorare ulteriormente la comprensione dei sistemi di IA, quindi rendere comprensibili ad un umano i meccanismi di funzionamento che altrimenti rimarrebbero 'nascosti', senza però ledere le performance dei sistemi. In questo modo si può garantire un corretto funzionamento delle IA, migliorando la safety, la security, la trasparenza, e la privacy.

3.1 Storia dei metodi di XAI: le tre generazioni

Secondo Mueller et al. [18] i metodi di XAI che sono stati pubblicati vanno storicamente divisi in tre generazioni. La prima generazione potrebbe essere ulteriormente divisa in due. Inizialmente sono stati pubblicati i cosiddetti *expert systems* di prima generazione che sono stati pubblicati agli inizi del 1970. Tali sistemi servivano per chiedere all'utente delle informazioni (come ad esempo i risultati di analisi del sangue) e fornivano in risposta delle raccomandazioni, o delle diagnosi. Essi fornivano rappresentazioni della conoscenza sotto forma di regole o reti di relazioni. Ma gli expert system non hanno poi ottenuto i benefici sperati. La prima generazione degli expert system, all'incirca, è durata una decina di anni fino alla nascita degli *explanation system* di prima generazione.

Questi sistemi riuscivano a fornire una spiegazione semplice che esprimeva le regole utilizzate per prendere una decisione. In generale, poiché la conoscenza e l'esperienza venivano rappresentate in termini di regole, tali regole potevano essere dotate di descrizioni in linguaggio naturale, che venivano sfruttate per fornire le risposte. Un esempio è quello fornito dal sistema XPLAIN (1985). Questo sistema dopo aver ricevuto le informazioni sulle analisi del sangue di un paziente poteva chiedere "Quali sono i livelli di calcio?",

l'utente avrebbe poi potuto chiedere "Perché?"; il sistema avrebbe quindi risposto "Il mio obiettivo è quello di iniziare la terapia, che include il controllo delle sensibilità. Ora sto misurando la sensibilità al calcio". Un problema fondamentale dei sistemi di prima generazione era l'incapacità di fornire una spiegazione sul come venissero fatte le inferenze, in particolare non era possibile capire il perché una certa azione venisse fatta o suggerita all'utente. Alla pubblicazione di questi metodi le problematiche citate comunque non erano state prese in considerazione, dato che avere delle risposte era già un grande passo avanti. Solo successivamente, quando questi sistemi iniziarono ad essere usati per l'istruzione, ci si pose il problema.

Già a partire dalla metà del 1980 si incominciano a produrre nuovi sistemi di seconda generazione. I ricercatori hanno iniziato a costruire sistemi che fossero in grado di tenere conto anche di un contesto, ad esempio imparando il dominio di applicazione o la storia personale dei pazienti. Non si limitavano a sintetizzare regole, ma si sforzavano di fornire spiegazioni comprensibili e rilevanti per l'utente. Questa generazione ha introdotto l'idea di spiegazioni contrastive, descrivendo perché alcune scelte non sono state fatte e confrontando diverse opzioni. Questo ha contribuito a una comprensione più approfondita dei processi decisionali. Hanno rappresentato un miglioramento significativo rispetto ai loro predecessori di prima generazione, concentrandosi sulla comprensibilità, sulla personalizzazione e sulla capacità di fornire spiegazioni più utili. Nonostante i miglioramenti, gli explanation system non fornivano delle vere e proprie spiegazioni ma sarebbe meglio dire delle giustificazioni.

Successivamente, per un ventennio (dal 1990 agli inizi degli anni 2010), il panorama dell'explainability ha visto un brusco calo nell'interesse, tanto da venire definito 'Explainability Winter'. Da dopo il 2010, grazie all'innovazione della tecnologia allo sviluppo di nuove tecniche di ML e DL il mondo dell'XAI si è risvegliato. Gli avanzamenti nelle IA e l'applicazione delle stesse a nuovi campi in settori in cui la sicurezza è determinante ha portato a nuove preoccupazioni sull'utilizzo non etico, su 'bias' indesiderati e la mancanza di trasparenza dei nuovi modelli. Molti nuovi articoli di ricerca si sono concentrati sulla visualizzazione, comparaione e comprensione del funzionamento delle reti neurali profonde (DNN), senza tralasciare anche il filone di ricerca che si è concentrato verso sistemi che producessero spiegazioni comprensibili anche ad utenti non sviluppatori.

3.2 Le classificazioni dei metodi

In letteratura sono presenti molti lavori che mirano a riunire sotto un'unica classificazione i lavori di XAI.

Nella pubblicazione di Arya et al. [5] viene proposto un nuovo toolkit *AI Explainability 360* che offre la possibilità di utilizzare diversi algoritmi di explainability ai sui propri modelli/dataset, insieme ad una tassonomia espressa sotto forma di decision tree. Tale tassonomia può essere navigata attraverso delle domande quando di attraversa un nodo, e i metodi di XAI consigliati si trovano sulle foglie. Alcune foglie, però, non elencano alcun metodo ma mostrano solamente un punto interrogativo.

Nella pubblicazione di Haar et al. [15] viene mostrata un'analisi nel panorama dell'XAI specificatamente per i metodi che operano su black-box di tipo CNN (Convolutional Neural Networks). L'analisi viene affrontata analizzando diverse categorie di metodi di XAI. Per ognuna di esse ne vengono illustrate le caratteristiche, un sommario di singole applicazioni su diversi domini e mostrando anche tecniche più innovative e meno diffuse.

Il lavoro di Guidotti et al. [14] fornisce un'ottima panoramica sul mondo dell'XAI, proponendo una tassonomia che identifica quali sono i problemi per l'apertura delle black-box generalizzando in quattro categorie. Tali categorie verranno poi riprese in seguito ed utilizzate nella nuova ontologia presentata in questo lavoro.

Speith [23] presenta un confronto tra diversi paper che hanno fornito una classificazione per i metodi di XAI. Nel suo articolo propone una nuova tassonomia ed inoltre, anche se solo descrivendone le caratteristiche generali e senza fornire una implementazione, due ulteriori proposte per superare le problematiche rilevate nel non avere una classificazione uniforme e consolidata nel panorama di ricerca.

Rimane però molto chiaro da tutti i lavori che non esiste una vera e propria classificazione universalmente accettata che copra tutti gli aspetti dell'argomento. Questo è dovuto non tanto dall'elevato numero di lavori che portano nuovi metodi, ma quanto alla grande diversità di funzionamento dei metodi.

3.3 Terminologia

Un passo fondamentale per procedere nell'analisi dell'argomento è quello di fare il punto sulla terminologia che verrà usata. Più di una volta è stata citata la 'spiegabilità' (explainability) dei sistemi ma senza aver ancora dato una definizione formale. La comunità scientifica, su questo tema, non ha ancora trovato un comune accordo. D'altronde una

definizione formale di cosa è 'spiegabile' rappresenta un nodo cruciale di intersezione tra le scienze cognitive e le scienze dell'informazione, che mira a indagare come gli esseri umani concepiscono e utilizzano le spiegazioni per comprendere i fenomeni e le azioni. Tutto ciò è necessario se si vuole creare e sviluppare sistemi di IA che siano non solo efficienti ma anche accettabili ed etici nella loro interazione con l'uomo.

Dai lavori di Graziani et al. [13] e Ali et al.[2] i quali mettono a confronto le definizioni fornite da molti articoli pubblicati, si evince che non esiste una accettabile terminologia che sia comune e consistente tra gli articoli. La mancanza di una definizione formale e unificata porta gli studiosi e i ricercatori ad interpretare e concentrarsi su aspetti diversi o persino divergenti del problema o del fenomeno in studio. Questa dispersione può comportare una serie di conseguenze negative:

- Ridotta coerenza e consolidamento delle conoscenze: potrebbe mancare una visione chiara e coerente dei progressi della comunità, rendendo difficile l'accumulo delle conoscenze consolidate.

- Spreco di risorse: gruppi di ricerca diversi potrebbero lavorare su uno stesso tema senza una collaborazione adeguata o un riferimento comune, duplicando inutilmente gli sforzi.

- Difficoltà di comunicazione: l'utilizzo di una terminologia non coerente può portare ad una comunicazione o uno scambio di informazioni incorretto e complicato, potenzialmente causando anche un ritardo nel progresso complessivo.

Detto ciò, risulta evidente che introdurre delle nuove definizioni, seppur giustificandole per renderle coerenti con il lavoro di ricerca, non apporterebbe alcun contributo. L'obiettivo dovrebbe essere l'ottenimento di una standardizzazione al fine di promuovere un'interpretazione uniforme. Dunque, si ritiene più opportuno riportare quelle che sono state già descritte da Ali et al. [2] avendo condotto un'analisi esaustiva delle similitudini e delle divergenze tra tali definizioni e quelle presenti in una serie di altri articoli.

Di seguito verranno elencate le principali definizioni che fanno parte dell'XAI. I primi due termini da evidenziare sono: *explainability* e *interpretability*. Secondo Adadi e Berrada [1] i due termini non possiedono una vera distinzione, tanto da usarli indistintamente durante la trattazione. Ciò nonostante viene aggiunto che "Explainable" è un termine più usato nel mondo dell'XAI mentre in ML viene di più utilizzato il termine "interpretable". Invece, Ali et al. [2] dà due definizioni ben distinte dei due termini e anche Graziani et al.

[13] specifica che in ambiti come le scienze sociali possono avere un diverso significato o essere pesate in maniera differente. Per cui si è preferito riportarle entrambe per evidenziare le differenze:

Definizione 3.1. (*Explainability*). Indica la capacità del modello di fornire una spiegazione sul *perché* ha fornito una certa risposta, in termini che possono essere comprensibili all'uomo. Più in seguito verranno mostrati i modelli di "Post-hoc Explainability" che indicheranno quegli algoritmi/metodi che serviranno per fornire spiegazioni alle decisioni dei modelli di AI [2].

Definizione 3.2. (*Interpretability*). L'interpretability è legata al capire *come* un modello fa le sue decisioni. Gli *intrinsic model* sono quegli algoritmi che spiegano in maniera comprensibile ad un umano il funzionamento di altri modelli [2].

Mentre gli expert system, citati in precedenza, attingevano ad una conoscenza esterna creata appositamente da esperti di settore, i moderni algoritmi di ML creano una conoscenza interna del mondo attraverso delle osservazioni sui dati, diventando la base delle predizioni. La complessità che hanno questi algoritmi è stata la chiave che ha permesso loro di ottenere una straordinaria abilità di predizione. L'altra faccia della medaglia è che questa complessità ha portato ad una difficile interpretazione del modello sottostante che nasconde, incapsulandola nei parametri, quella conoscenza che ha osservato. Facendo un esempio le reti neurali profonde (DNN) sono delle strutture multi strato non lineari, composte da diversi strati nascosti e con numerosi nodi per strato. Se una singola trasformazione lineare può essere interpretabile (osservando pesi, input e output), un insieme complesso di interazioni che si mescolano ad ogni livello diventa estremamente complesso, se non impossibile, da riuscire a comprendere nella sua interezza.

Ci sono poi altre definizioni importanti, tutte volte a migliorare la fiducia nei modelli di IA. Questi concetti sono spiegati di seguito.

Definizione 3.3. (*Trasparency*). La trasparency di un modello è ottenibile quando un intrinsic model (vedi Definizione 3.2) può generare delle explanations per esso. La trasparency è una caratteristica fondamentale per stabilire la qualità di un modello di IA [2].

Definizione 3.4. (*Fairness*). La fairness si riferisce all'abilità di un modello di prendere decisioni senza 'bias' ovvero senza pregiudizi o sbilanciamenti in favore o contro un certo sottoinsieme della popolazione [2]. Questo è uno degli aspetti cruciali per creare un buon

sistema di IA. Tale concetto è spesso menzionato nell'AiAct e nel quale sono state spese molte parole in merito.

Definizione 3.5. (*Robustness*). La robustness indica la capacità di un modello di non essere sensibile alle piccole variazioni di input. Un modello dovrebbe essere attentamente addestrato in modo da fornirgli una scarsa sensibilità alle variazioni, per garantire un corretto funzionamento in caso di incertezza [2].

Definizione 3.6. (*Satisfaction*). La satisfaction è una misura relativa ai sistemi di XAI, e indica il grado di miglioria che può apportare un sistema di ML/DL [2].

Definizione 3.7. (*Stability*). La stability è la misura della capacità di un modello di XAI a fornire una risposta comparabile per input simili [2].

Definizione 3.8. (*Responsibility*). La responsibility indica l'affidabilità di un modello che viene incrementata se esso rispetta le norme di socialità, etica, morale, correttezza e trasparenza [2].

Risulta evidente che queste definizioni non forniscano indicazioni per creare vere misure oggettive o su come utilizzarle. D'altronde non esiste una maniera per misurare l'etica o la morale.

4 ExOn: una nuova ontologia

Come illustrato in precedenza, le tecniche di explainability mirano a rispondere alle domande sul "come" e sul "perché" una IA ha fornito le proprie risposte e decisioni. Per fornire un servizio il più possibile corretto, sia in termini di correttezza semantica delle risposte che correttezza etica e morale. Porsi tali domande è un passo fondamentale per la pubblicazione di questi strumenti intelligenti, perché servono a spiegare i funzionamenti "nascosti" che le black-box possiedono.

A tal proposito si vuole presentare uno strumento in grado di poter fornire una visione più chiara del panorama dell'XAI grazie ad una ontologia pensata per essere resiliente ai nuovi progressi nel campo. ExOn (Explainability Ontology) è una nuova ontologia che permetterà di classificare secondo le più recenti categorizzazioni pubblicate nello stato dell'arte le pubblicazioni sui metodi di explainability, ed inoltre, fornirà una prospettiva sulle connessioni e sulle evoluzioni della ricerca grazie ad alcune proprietà esprimibili utilizzando le formalizzazioni messe a disposizione dalle ontologie.

Attraverso l'uso di questa ontologia, gli utenti avranno la possibilità di filtrare le pubblicazioni secondo la personale necessità, oppure di valutare quali sistemi di explainability hanno avuto più successo in base alla specifica classe. Mentre i ricercatori possono ottenere una migliore panoramica informativa nell'ambito.

4.1 Le ontologie

Nella filosofia il termine *ontologia* indica "lo studio dell'essere in quanto tale, nonché delle sue categorie fondamentali". In altre discipline, come nel software engineering o nell'IA, è definita come "una specificazione formale esplicita di una concettualizzazione condivisa". [11].

Le ontologie possono essere utili per condividere la conoscenza. In alcuni ambiti, come il dominio dei dispositivi elettronici, una ontologia avrebbe al suo interno termini specifici (transistori, diodi, ...), ma anche termini più generali (funzioni, processi, ...). L'ontologia cattura la struttura concettuale intrinseca del dominio. Per costruire un linguaggio di rappresentazione della conoscenza basandoci sull'analisi, è necessario associare i termini con i concetti e le relazioni nell'ontologia e creare una sintassi per codificare la conoscenza. Possiamo quindi condividere questo linguaggio con coloro i quali necessitano di conoscere il dominio [6].

Il linguaggio che è stato utilizzato per rappresentare la conoscenza è il linguaggio OWL 2 creato dal W3C [9]. I concetti base per rappresentare la conoscenza con questo linguaggio sono: gli assiomi, le entità e le espressioni.

- Gli *assiomi* sono le affermazioni espresse da un'ontologia OWL. (Es. "John è un maschio ed è sposato con Mary")
- Le *entità* sono gli elementi usati per riferirsi agli oggetti del mondo reale. (Es. John, Mary, maschio, sposato)
- Le *espressioni* sono combinazioni di entità semplici per formare entità più complesse.

È possibile dividere le entità in ulteriori elementi: individui, classi e proprietà.

- Una *classe* è un'astrazione di un insieme di oggetti con delle caratteristiche in comune.
- Un *individuo* è un oggetto identificabile della realtà.
- Le *proprietà* sono le relazioni che intercorrono tra le entità. Esse sono ulteriormente divisibili in: object property, data properties e annotation properties.
  - Le *object property* esprimono una relazione tra 2 individui, tra 2 classi, o tra individui e classi (ad esempio lega una persona alla propria sposa).
  - Le *data property* assegnano un valore ad un oggetto (ad esempio l'età ad una persona).
  - Le *annotation property* sono delle informazioni aggiuntive sugli assiomi utili in fase di navigazione dell'ontologia (ad esempio esplicita l'autore di un certo assioma).

Quindi, le ontologie hanno la capacità di mostrare relazioni semantiche tra concetti e attributi, essendo una struttura gerarchica di termini per descrivere un dominio, e può essere usata come base di conoscenza (knowledge base). Infine, un'altra caratteristica è la possibilità di riusare ontologie già presenti o collegare due ontologie. In questo modo si può estendere e migliorare la conoscenza sfruttando una struttura formale ben definita.

Le ontologie sono dunque degli strumenti molto potenti perchè offrono la possibilità di creare strutture gerarchiche che possono rappresentare conoscenze complesse e facilitano la comprensione delle relazioni tra concetti e dati, hanno una chiara rappresentazione concettuale grazie alla standardizzazione e possono anche permettere il ragionamento automatico.

4.2 Le classi di ExOn

La classe centrale che l'ontologia presenta è la classe *Publications*. Una sua istanza deve essere intesa, nello scopo di questa ontologia, come una pubblicazione scientifica che mostri una nuova tecnica nell'ambito dell'XAI. L'identificativo che caratterizza ogni pubblicazione sarà una stringa composta dall'URI dell'ontologia `http://www.semanticweb.org/alexm/ontologies/2023/8/ExOn` seguita da un asterisco (`#`) e dal DOI della pubblicazione. Insieme ad essa, è presente un insieme di altre 5 classi, e relative sottoclassi, connesse alla classe *Publications* attraverso delle proprietà. Lo schema di base è mostrato in Figura 2. Le classi di primo livello vengono descritte nella Tabella 1. In essa, l'ultima colonna presente riporta per ogni classe il riferimento corrispondente alle tabelle che elencano le relative sottoclassi con l'eventuale descrizione approfondita.

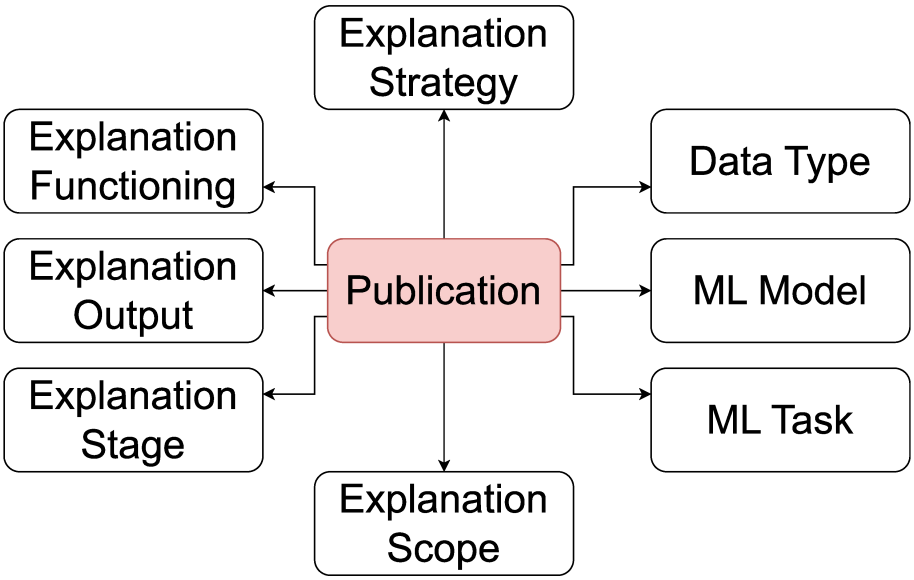

Figura 2: Classi di primo livello di ExOn. Il rettangolo rosso di Publication sta a significare che sono le istanze di Publication ad essere collegate con le proprietà alle classi esterne.

4.3 Le proprietà dell'ontologia

Come accennato, in ExOn sono state definite delle proprietà, o relazioni, che collegano le istanze della classe Publication alle altre classi. La classificazione delle istanze di Publica-

Classe	**Descrizione**	**Sottoclassi**
Publication	La classe fondamentale che viene collegata a tutte le altre classi. Rappresenta il modello di Explainability presentato nella pubblicazione.	No
ML Model	Rappresenta il modello black-box sul quale l'explanation viene effettuata.	Tab. 2
ML Task	Rappresenta il task che il modello black-box porta a termine. Viene ereditato e incrementato l'elenco presentato da Nauta et al. [20]	Tab. 3
Data Type	Rappresenta il tipo di dati con cui il modello di explainability lavora.	Tab. 4
Expl. Strategy	Rappresenta il 'problema' che il modello di explainability vuole risolvere. Tale classificazione è un leggero adattamento della classificazione riportata da Guidotti et al. [14] con il nome di *Open the black box problem*.	Tab. 5
Expl. Functioning	Rappresenta il funzionamento del modello di explainability, ovvero il modo in cui si estraggono informazioni dai modelli di ML. Tale classificazione è stata descritta da Speith [23] con il nome di *Functioning-based approach* e poi utilizzata nella tassonomia finale.	Tab. 6
Expl. Stage	Distingue tra le tecniche applicate a priori o a posteriori. Tale classificazione è stata descritta da Speith [23], ma rappresenta una classificazione comune nei review paper.	Tab. 7
Expl. Output	Rappresenta l'output che il metodo di explainability fornisce come spiegazione. [20]	Tab. 8
Expl. Scope	Rappresenta la generalità con la quale il metodo di explainability effettua le spiegazioni. Anche questa rappresenta una classificazione comune tra i review paper.	Tab. 9

Tabella 1: Classi di primo livello di ExOn

ML Model	**Note**
Any Model	Classe necessaria per i modelli XAI model-agnostic
Multi Layer Neural Network	-
Convolutional Neural Network	-
Recurrent Neural Network	-
Support Vector Machine	-
Ensembles and Multiple Classifier Systems	-
Other	Altri modelli

Tabella 2: Sottoclassi di ML Model

ML Task	**Note**
Any Task	Classe necessaria per i modelli XAI model-agnostic. [20]
Classification	[20]
Regression	[20]
Policy Learning	[20]
Binary Classification	Caso particolare di classificazione a due classi.
Clustering	-
Text Processing	-
Natural Language Processing	-
Other	[20]

Tabella 3: Sottoclassi di ML Task

Data Type	**Note**
Any Data Type	Classe necessaria per i modelli XAI data-agnostic.
Graph	-
Image	-
Tabular	Comprende anche dati strutturati.
Text	-
Time Series	-
User Item Matrix	Matrici utenti-oggetti usate nei recomandation systems.
Other	-

Tabella 4: Sottoclassi di Data Type
Tutti i Data Type elencati fanno riferimento alla classificazione descritta da Nauta et al. alla quale si rimanda per approfondimenti [20].

Expl. Strategy	**Note**
Model Explanation	È il problema della creazione di un modello di explainability globale che fornisca delle spiegazioni.
Model Inspection	È il problema di fornire una rappresentazione di quale proprietà della black box in forma visuale o testuale.
Outcome Explanation	È il problema della creazione di un modello di explainability locale che fornisca una spiegazione ad una particolare istanza.
Transparent Box Design	Consiste nel creare un modello localmente o globalmente interpretabile.

Tabella 5: Sottoclassi di Expl. Strategy
Queste classi sono riprese da Guidotti et al. [14] dove venivano presentate come l'"*Open the black box problems*".

Functioning	**Note**
Architecture Modification	Si modifica la struttura della black-box per ottenere una architettura semplificata più interpretabile.
Examples	Si cercano degli elementi nel dataset di training che sono rappresentativi delle classi, ad esempio per la classificazione.
Meta-Explanations	Si utilizzano diversi modelli di explaninability per formare una spiegazione finale combinata.
Perturbations	Si perturbano gli input dei modelli black box per valutare la modifica nell'output.
Structure Leveraging	Si vanno ad esaminare specifiche proprietà del modello black box, ad esempio osservando i gradienti.

Tabella 6: Sottoclassi di Expl. Functioning
Queste classi sono riprese da Samek and Muller [22] e Arrieta et al. [3] e messe insieme nella tassonomia di Speith [23]

Expl. Stage	**Note**
Ante-Hoc	Indica le tecniche per la creazione di metodi white-box interpretabili a priori.
Post-Hoc	Indica le tecniche per ottenere un'interpretazione dei metodi black-box. Ha un'ulteriore livello di sottoclassi che sono: • Model-agnostic: indica i metodi di explainability che funzionano indistintamente dal modello black-box che vanno a spiegare. • Model-specific: indica i metodi di explainability specifici per un modello black-box.

Tabella 7: Sottoclassi di Expl. Stage

Expl. Output	**Note**
Binary Feature Importance	Valore binario da associare alle feature in input. Ad esempio patch di immagini, segmentation o bounding box.
Decision Rules	Regole logiche che includono decision set, anchors e decision table
Decision Tree	Albero decisionale con condizioni su ogni nodo. Un esempio è ProtoTree.
Feature Importance	Valori non binari associati alle feature in ingresso che indicano feature relevance, attribution o la contribution. Esempi possono essere i metodi SHAP e gli importance score di LIME.
Feature Plot	Figure che mostrano le relazioni o le interazioni tra feature o tra feature e l'output. Esempi possono essere Partial Dependance Plot, Individual Conditional Expectation plot o Feature Auditing.
Graph	Grafi con nodi e archi che mostrano graficamente la struttura. Esempi possono essere Abtract Policy Graph, Knowledge graph, Flow graph, e Automi a Stato Finito.
Heatmap	Mappe di almeno 2 dimensoni che evidenziano visualmente feature attribution, activation, sensitivity, attention o saliency. Esempi possono essere attention maps, perturbation masks o Layer-Wise Relevance Propagation.
Model	Un modello white-box intrinsicamente interpretabile, che funge da spiegazione al modello black-box. Esempi possono essere scoring sheet o linear regression.
Prototypes	Elementi chiave del dataset molto rappresentativi. Include anche concepts, istanze influenti per il training, prototypical parts, nearest neighbors e criticism.
Text	Spiegazioni testuali attraverso linguaggio naturale.
Other	Altri tipi di explanation che non appartengono a nessun'altra categoria.

Tabella 8: Sottoclassi di Expl. Output
Tutti gli output descritti fanno riferimento alla classificazione descritta da Nauta et al. alla quale si rimanda per approfondimenti. [20]

Expl. Scope	Note
Global	Sono quei metodi che affrontano il problema di creare una spiegazione sul funzionamento generico della black-box. Inoltre, i metodi ante-hoc dovrebbero riferirsi a questa classe.
Local	Sono quei metodi che affrontano il problema di creare una spiegazione sul funzionamento di una singola istanza della black-box.

Tabella 9: Sottoclassi di Expl. Scope

tion la si ottiene esprimendo le relazioni che tali istanze hanno con le classi descritte nella sezione precedente.

4.3.1 Le object property

Secondo la sintassi OWL, una relazione tra una istanza ed una classe (o un'altra istanza) viene definita *object property*. In Figura 3 viene mostrato graficamente l'insieme delle object property che collegano l'istanza di Publication alle altre classi (istanze).

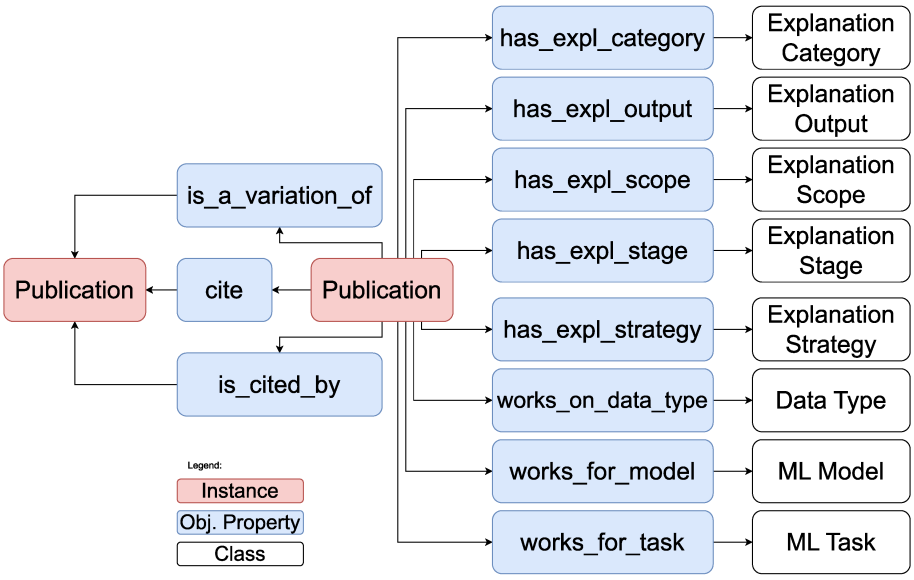

Figura 3: Le object property definite in ExOn collegate alle relative risorse.

Un elenco delle object property presenti in ExOn è mostrato nella Tabella 10, dove vengono specificate anche delle caratteristiche aggiuntive.

Object property	**Note**
cite	È una proprietà la lega due pubblicazioni. È definita come la proprietà inversa (*owl:inverse_of*) della proprietà is_cited_by.
has_expl_category	È la proprietà che lega una pubblicazione all'Expl. Category che il metodo in essa contenuto utilizza.
has_expl_output	È la proprietà che lega una pubblicazione all'Expl. Output che il metodo in essa contenuto fornisce.
has_expl_scope	È la proprietà che lega una pubblicazione all'Expl. Scope che il metodo in essa contenuto utilizza.
has_expl_stage	È la proprietà che lega una pubblicazione all'Expl. Stage che il metodo in essa contenuto utilizza.
has_expl_strategy	È la proprietà che lega una pubblicazione all'Expl. Strategy che il metodo in essa contenuto utilizza.
is_a_variation_of	È la proprietà che lega una pubblicazione alla pubblicazione con la variante originale del metodo di XAI.
is_cited_by	È la proprietà inversa di *cite*.
works_for_data_type	È la proprietà che lega una pubblicazione alle tipologie di dati che il metodo in essa contenuto utilizza.
works_for_model	È la proprietà che lega una pubblicazione al modello black-box che il metodo di XAI spiega.
works_for_task	È la proprietà che lega una pubblicazione alla tipologia di task del modello black-box che il metodo di XAI spiega.

Tabella 10: Le object property di ExOn
Da intendere che il dominio delle proprietà è sempre una istanza di Publication, mentre il codominio è visualizzabile nella Figura 3

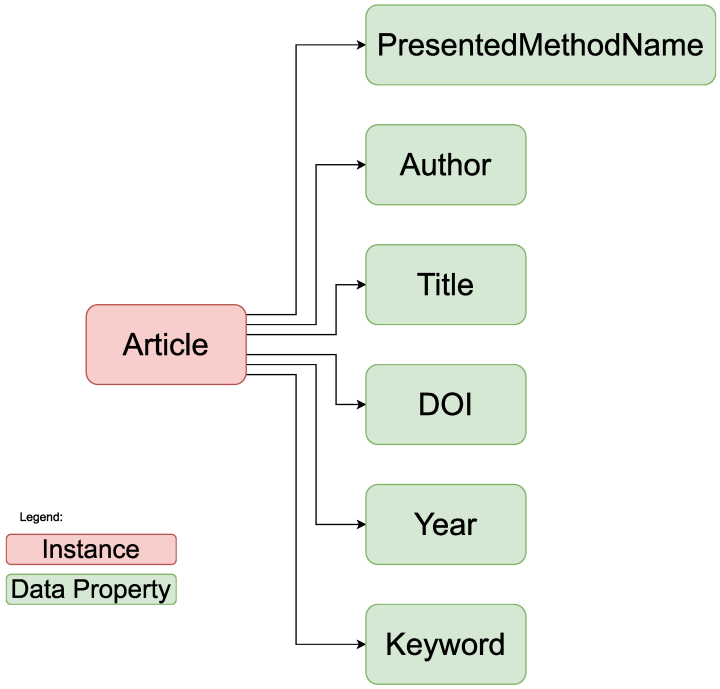

Figura 4: Le data property definite in ExOn. Il tipo di risorsa base (stringa o intero) è stato omesso.

Un caso particolare di object property sono la **cite**, **is_cited_by** e **is_a_variation_of** poiché esse collegano le publication ad altre publication. Queste relazioni ricorsive permettono di ottenere un collegamento tra le pubblicazioni che può essere grado di evidenziare i collegamenti e le influenze che ci sono state tra le pubblicazioni, identificando le correnti di ricerca.

4.3.2 Le data property

Oltre alle object property, in ExOn sono stati definiti anche altri tipi di relazioni che legano le istanze di Publication a dei valori di tipo base. Tali relazioni vengono definite *data property*. Queste proprietà permettono di inserire informazioni aggiuntive per ogni pubblicazione, senza la necessità di modellare formalmente i valori. In questo modo sarà comunque possibile effettuare ricerche tra le istanze delle pubblicazioni secondo le data property. In Figura 4 è mostrato uno schema grafico delle data property di ExOn.

Alcune data property sono state definite con una restrizione permessa dalla sintassi

Data property	OWL data type	OWL functional
Author	rdf:PlainLiteral	×
DOI	rdf:PlainLiteral	✓
Keyword	rdf:PlainLiteral	×
PresentedMethodName	rdf:PlainLiteral	×
Title	rdf:PlainLiteral	✓
Year	xsd:integer	✓

Tabella 11: Le data property di ExOn
Da intendere che il dominio delle data property è sempre Publication.

OWL; tale restrizione è la *owl:functional* la quale serve ad indicare che per ogni istanza è al massimo possibile esprimere un solo valore per quella proprietà. La Tabella 11 comprende l'elenco completo delle data property insieme alla specifica della restizione applicata ad ogni proprietà.

4.4 Strumenti utilizzati

Sono disponibili molti applicativi in grado di fornire supporto per la creazione di nuove ontologie. Il programma che è stato scelto e con il quale è stata creata ExOn è Protégé (v 5.5.0) [19], il quale mette a disposizione un'applicativo con interfaccia desktop che semplifica la produzione delle ontologie. La sua natura open source, la possibilità di gestire in maniera efficiente ontologie di grandi dimensioni e il suo utilizzo in organizzazioni accademiche, governative e private in ambiti che spaziano tra la medicina e il commercio online lo rendono un'ottima scelta. Infine, Protégé supporta le specifiche dell'Ontology Web Language 2 (OWL 2) e RDF del World Wide Web Consortium (W3C).

All'interno di Protégé è possibile installare dei plug-in i quali servono ad estendere le funzionalità del programma. I plugin utilizzati sono:

- SPARQL (v 6.0.0) che permette di effettuare query verso dati che seguono le specifiche RDF come, ovviamente, le ontologie in linguaggio OWL.
- OntoGraf (v 2.0.3) che è stato utile durante lo svoluppo per navigare interattivamente le relazioni dell'ontologia.

4.5 Esempi di ricerca

Un'ontologia scritta in linguaggio OWL può equivalentemente essere espressa come un grafo RDF. Il W3C ha messo a disposizione il linguaggio di interrogazione SPARQL che permette di eseguire query di interrogazioni che restituiscono dati che seguono le specifiche RDF. Essendo i grafi costituiti da triple soggetto-predicato-oggetto, ciò può essere visto come un analogo di alcuni database NoSQL key-value come mongodb.

Come primo esempio mostriamo il codice per poter selezionare tutte le istanze che sono di classe Pubblication.

```
PREFIX ExOn: <http://www.semanticweb.org/alexm/ontologies/2023/8/ExOn#>
SELECT ?x
WHERE {
    ?x a ExOn:Article
}
```

Una caratteristica che è stata menzionata più volte è la capacità dell'ontologia di tenere traccia dell'andamento degli sviluppi della ricerca. Un modo per farlo è sfruttare la object property is_a_variation_of descritta precedentemente. Nell'esempio successivo è possibile estrarre dall'ontologia i metodi con più citazioni che sono variazioni di una pubblicazione.

```
SELECT ?Pub ?Title (COUNT(?cit) as ?count_cit)
WHERE {
    ?A a ExOn:Publication;
        ExOn:PresentedMethodName "LIME".
    ?Pub a ExOn:Publication;
        ExOn:is_a_variation_of ?A;
        ExOn:Title ?Title.
    optional{
        ?cit a ExOn:Publication;
            ExOn:cite ?Pub
    }
}
GROUP BY ?Pub ?Title
ORDER BY ?count_cit
```

Il codice va letto in questo modo: "Data una pubblicazione A che presenta un metodo di nome "LIME" estrarre tutte le pubblicazioni che sono una variazione di A mostrandone il titolo e il numero di citazioni che hanno ricevuto e porli in ordine decrescente rispetto al numero di citazioni". Ovviamente è possibile scendere di 'livello' e vedere per ogni elemento risultante quali sono le variazioni che hanno avuto più successo, o che non sono mai state modificate.

Questo dà il senso delle potenzialità di ExOn, che non si riduce ad un semplice archivio, ma ha le capacità per essere utilizzato in maniera automatica ed intelligente per estrarre della conoscenza.

5 Discussione

Dopo aver elencato le caratteristiche di ExOn, verranno ora discussi alcuni vantaggi che l'utilizzo di una ontologia come classificazione può portare, insieme a dei confronti rispetto ad alcune classificazioni esistenti mostrando come una ontologia ideata in questo modo sia un miglioramento dello stato dell'arte.

Già è stato citato molte volte il fatto che un'ontologia, proprio per sua natura, sia una descrizione formale della realtà quindi può ammettere una espansione teoricamente indefinita se correttamente modellata da principio. Le classificazioni o le tassonomie sono più rigide sotto questo punto di vista.

Un altro aspetto è che tramite le ontologie si ha la possibilità di esprimere direttamente la struttura della conoscenza e quindi abilitare le macchine ad effettuare elaborazioni automatiche sulla conoscenza stessa e quindi dedurre nuove informazioni. Il web semantico sotto questo punto di vista è da molti anni un obiettivo del W3C, nel tentativo di creare sistemi che interagiscano tra di loro in sicurezza. Per farlo sono state sviluppate le tecnologie come l'RDF, il linguaggio OWL, SPARQL ed altre.

Nei successivi paragrafi verrà messa a confronto ExOn con altre classificazioni presenti nello stato dell'arte, valutando le migliorie apportate allo stato dell'arte.

5.1 Confronto con Arrieta et al.

Il primo confronto che affrontiamo è tra ExOn e un lavoro molto influente in questo ambito ovvero la tassonomia di Arrieta et al. [3]. In Figura 5 è possibile osservare lo schema proposto da Arrieta et al. dove da sinistra a destra la classificazione entra sempre di più nello specifico del metodo fino ad arrivare all'ultimo livello nel quale sono presenti riferimenti ad articoli di metodi di XAI presentati in letteratura. Da notare come i colori che contraddistinguono i riferimenti stanno ad indicare la tipologia di dati trattati (blu, verde e rosso corrispondono rispettivamente a immagini, testo e dati tabulari).

Se un utente volesse fare una ricerca su questa gerarchia, esso sarà costretto a partire dalla generalità del metodo (Expl. Stage in ExOn), poi passare per la specificità del metodo (Expl. Scope in ExOn), il metodo di ML da spiegare, il funzionamento del metodo di explainability (Expl. Functioning in ExOn) ed infine il tipo di output della spiegazione. Solo a questo punto può apprendere gli articoli che soddisfano le proprie necessità.

ExOn, rispetto ad Arrieta, porta tutte le dimensioni al primo livello, invece che usare una gerarchia. In questo modo si evita non solo la ricerca secondo lo schema gerarchico, ma

anche il possibile problema che nel futuro possa essere pubblicato un nuovo metodo il quale non potrà trovare spazio nella classificazione. In quest'ultimo caso in ExOn basterebbe creare una nuova istanza di Publication e attribuirle tutte le proprietà necessarie allo scopo di categorizzarla, e se si dovesse ritenere opportuno estendere l'ontologia.

Infine si può notare come lo schema riporti le tecniche di exlpainability (equivalente di Explanation Functioning in ExOn), diverse volte per tipologia di metodo di ML. Questo porta ad avere una ripetizione sullo schema che rende la ricerca degli articoli con quella specifica tecnica e indipendentemente dal metodo di ML molto complessa.

5.2 Confronto con Speith

Un altro confronto che viene mostrato è con il lavoro di Speith [23], il quale dopo uno studio di diverse classificazioni nello stato dell'arte propone di riunire le idee comuni sotto un'unica nuova tassonomia. Lo schema presentato è visibile in Figura 6.

L'approccio che è stato utilizzato è quello di inserire tutte le dimensioni di categorizzazioni in modo da essere direttamente collegate con il metodo di explainability. Questo porta il vantaggio che ogni classe risulta indipendente dalle altre, con il risultato di garantire che tutte le combinazioni siano eventualmente possibili.

Rispetto a Speith, ExOn migliora la classificazione dato che oltre ad aumentare il numero delle sottoclassi definite dando così maggior possibilità di discriminazione, si tiene anche conto dei metadati delle pubblicazioni di XAI dando la possibilità all'utente di osservare l'andamento della ricerca come già discusso precedentemente. Nonostante ciò risulta evidente come il lavoro di Speith sia stato una base di partenza per ExOn, tanto che alcune nomenclature delle classi sono state scelte a partire da quelle descritte su questo lavoro.

5.3 Confronto con Nauta et al.

L'ultimo confronto che viene discusso è quello con la classificazione di Nauta et al. [20]. Lo schema della classificazione, che riporta le categorie di primo livello, viene riportato in Figura 7.

ExOn ha preso spunto dal lavoro di Nauta et al., tanto da aver riportato con lo stesso significato alcune categorie. Molto utile è stato il lavoro sulla categorizzazione dei Type of Explanation (Explanation Output in ExOn), poiché risulta un elenco sufficientemente completo e generale per rappresentare tutte le tipologie di spiegazione.

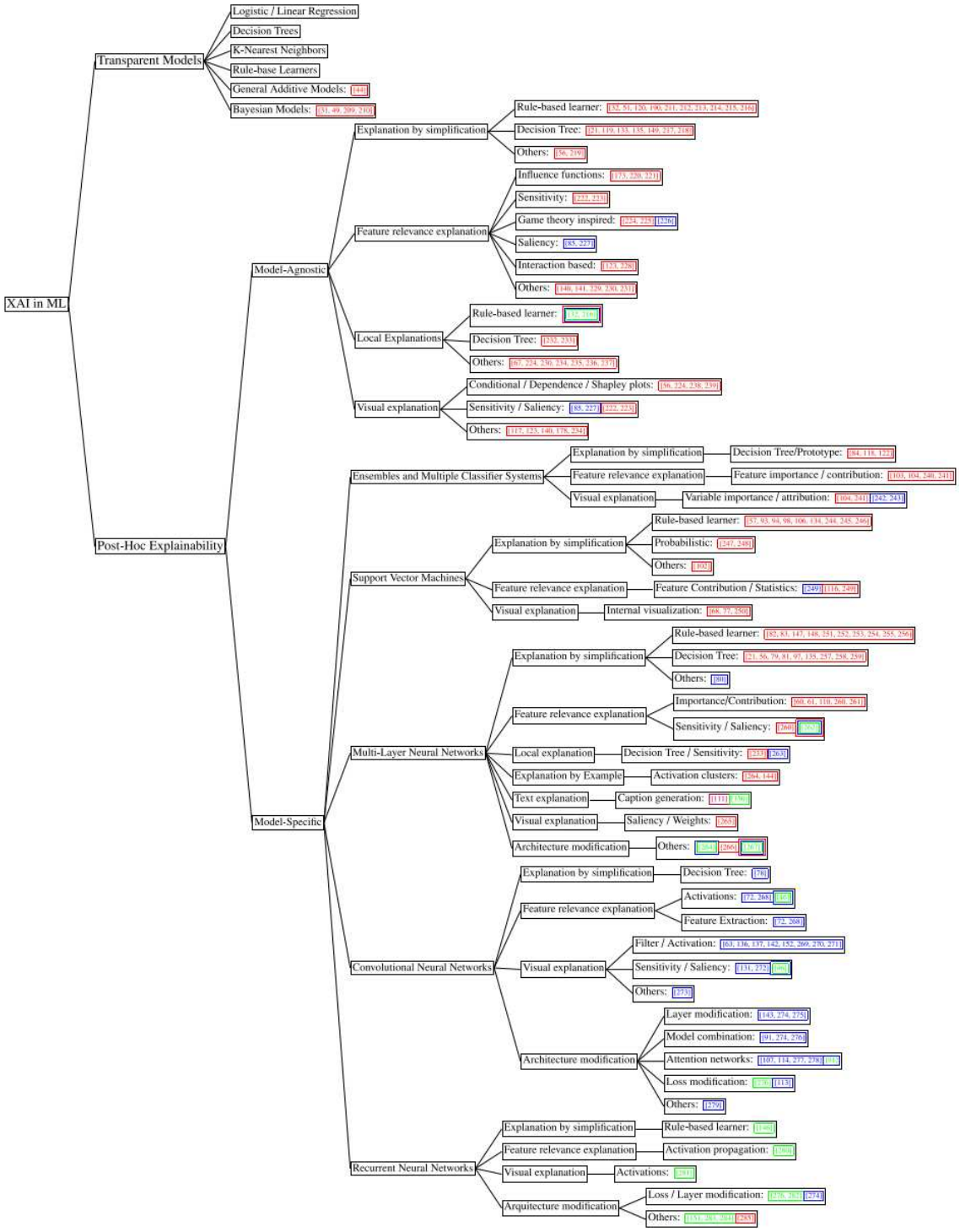

Figura 5: Classificazione di Arrieta et al. [3]

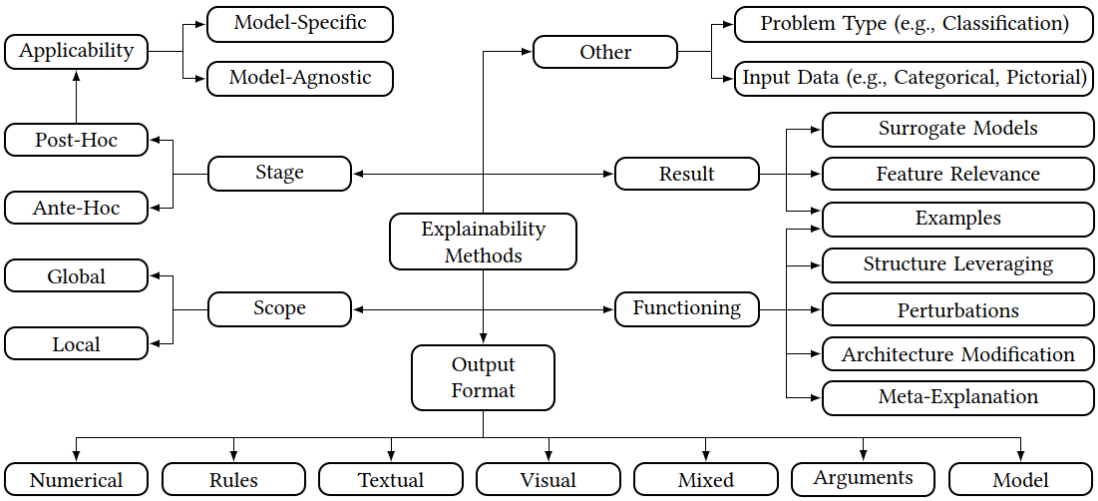

Figura 6: Classificazione di Speith [23]

Si può però notare la mancanza di una classe base, ovvero quella definita in ExOn con il nome di Explanation Scope, che include le due categorie local e global. Distinguere i metodi tra locali e globali è molto comune nella letteratura, tanto che questo fatto viene menzionato da Nauta et al., ma non viene inserito tra le categorie di classificazione. Essendo uno dei pochi punti che accomuna tutte le classificazioni, è stato ritenuto fondamentale inserirlo in ExOn.

Nella loro pubblicazione è presente anche un riferimento al database di articoli sui metodi per l'XAI [21] che è stato creato dagli autori a supporto del loro studio. In questo database sono stati inseriti e classificati più di 300 articoli, e dal quale sono partiti per elencare le 12 proprietà che una spiegazione dovrebbe avere.

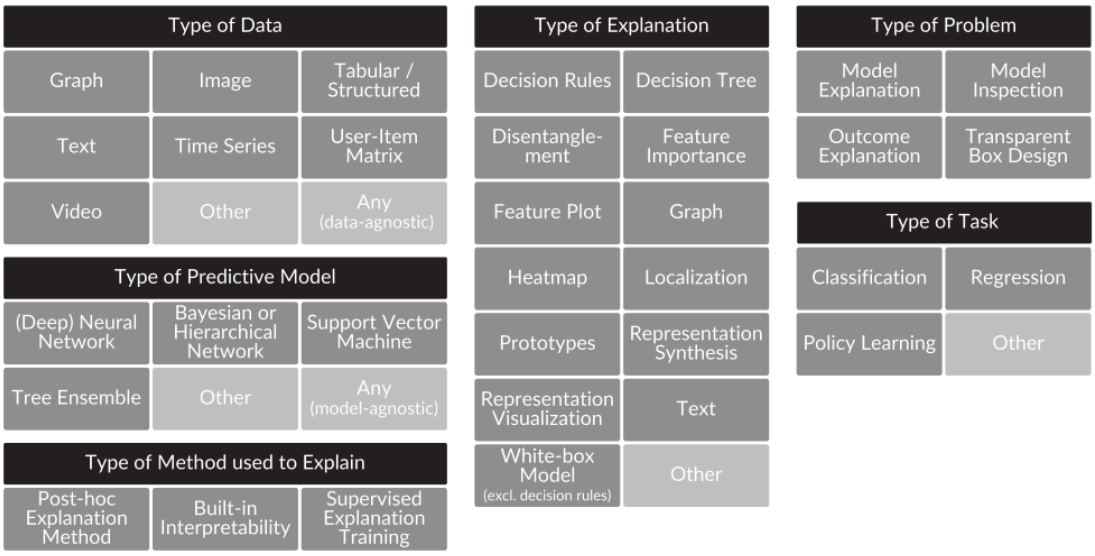

Figura 7: Classificazione di Nauta et al. [20]

6 Conclusioni e sviluppi futuri

In questa dissertazione si è parlato del mondo dell'eXplainable Artificial Intelligence (XAI), della sua storia, delle motivazioni della sua esistenza, e di come i significativi avanzamenti dell'intelligenza artificiale degli ultimi anni hanno evidenziato questioni di rilevanza non trascurabile come: preoccupazione riguardo all'utilizzo improprio delle IA, i potenziali problemi di privacy o sicurezza se non correttamente controllati, o bias nascosti che potrebbero portare a delle discriminazioni in fase di utilizzo dei nuovi sistemi. La nuova normativa AIAct che è stata presentata in Unione Europea mette in luce queste ed altre preoccupazioni rendendo necessaria la ricerca nel settore dell'XAI.

Per questo viene presentata la nuova ontologia ExOn (Explainability Ontology), che permette di formalizzare il concetto di pubblicazione nella quale viene descritta una tecnica di XAI.

Dai confronti con le classificazioni in letteratura risulta che ExOn migliora lo stato dell'arte portando il concetto di classificazione da una semplice tassonomia dei metodi ad una formalizzazione espressa sotto forma di ontologia. Inoltre, le categorie utilizzate sono derivate tenendo in considerazione le categorizzazioni più recenti in modo da cercare di rappresentare ogni aspetto dei metodi.

Infine, risulta che la classificazione presentata permette di evidenziare i collegamenti e le influenze che ci sono tra le pubblicazioni, grazie alla formalizzazione dei concetti di 'citazione' e di 'metodo derivato da' migliorando l'aspetto della tracciabilità della ricerca.

6.1 Sviluppi futuri

Come accennato nella sezione 5, una ontologia può ammettere un'espansione teoricamente indefinita. Considerato ciò, sarebbe possibile in futuro espandere ExOn ad esempio formalizzando il concetto di autore di una pubblicazione, attualmente inserito come una data property. Si potrebbe quindi estendere l'esplorazione del mondo delle pubblicazioni oltre che dal punto di vista del 'documento' anche da quello degli autori. Per cui ci sarebbe la possibilità di osservare l'influenza che i singoli autori hanno avuto sugli altri. Grazie alla flessibilità delle ontologie, per farlo potrebbe non essere necessario elaborare una nuova formalizzazione ma sfruttare altre ontologie, o parti di esse, già presenti nello stato dell'arte.

Un altro possibile aspetto sul quale si può apportare un miglioramento sta nella formalizzazione dei modelli di machine learning, sui quale si effettuano le spiegazioni. In particolare Klampanos et al. [16] ha presentato ANNETT-O, un'ontologia che è in grado di modellare gli aspetti riguardanti la topologia, il training e la valutazione delle deep neural network in qualsiasi configurazione. L'integrazione di ANNETT-O ad ExOn trova spazio proprio nella sottoclasse ML Models, che potrebbe essere adattata in futuro per accogliere ANNETT-O. In questo modo si può aumentare ancora di più la formalizzazione di questo dominio.

Infine, sarebbe da valutare la possibilità di costruire un applicativo ad interfaccia grafica che sfrutti ExOn come knowledge base ma migliori la sua navigazione anche ad utenti che non hanno familiarità con il linguaggio di interrogazione SPARQL, seppur offrendo la possibilità di mantenere le caratteristiche di formalità di un'ontologia.

7 Riferimenti bibliografici

[1] Amina Adadi e Mohammed Berrada. «Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)». In: *IEEE access* 6 (2018), pp. 52138–52160.

[2] Sajid Ali et al. «Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence». In: *Information Fusion* 99 (2023), p. 101805.

[3] Alejandro Barredo Arrieta et al. «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI». In: *Information fusion* 58 (2020), pp. 82–115.

[4] *Artificial Intelligence Act*. Awaiting Parliament's position in 1st reading. European Commission, 21 apr. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[5] Vijay Arya et al. «One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques». In: *arXiv preprint arXiv:1909.03012* (2019).

[6] Balakrishnan Chandrasekaran, John R Josephson e V Richard Benjamins. «What are ontologies, and why do we need them?» In: *IEEE Intelligent Systems and their applications* 14.1 (1999), pp. 20–26.

[7] *Che cos'è l'intelligenza artificiale?* Parlamento Europeo, 28 giu. 2023. URL: https://www.europarl.europa.eu/news/it/headlines/society/20200827STO85804/che-cos-e-l-intelligenza-artificiale-e-come-viene-usata.

[8] European Commission, Directorate-General for Research e Innovation. *Horizon Europe, budget – Horizon Europe - the most ambitious EU research and innovation programme ever*. Publications Office of the European Union, 2021. DOI: doi/10.2777/202859.

[9] W3C Commission. Dic. 2012. URL: https://www.w3.org/TR/owl2-primer/.

[10] *EU to invest €180 million in cutting-edge digital technologies and research*. European Commission, ago. 2023. URL: https://digital-strategy.ec.europa.eu/en/activities/digital-technologies-and-research.

[11] Jérôme Euzenat et al. «State of the art on ontology alignment». In: *Knowledge Web Deliverable D* 2 (2004), pp. 2–3.

[12] Steven Feldstein. *The Global Expansion of AI Surveillance*. Carnegie, 17 set. 2019. URL: https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

[13] Mara Graziani et al. «A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences». In: *Artificial intelligence review* 56.4 (2023), pp. 3473–3504.

[14] Riccardo Guidotti et al. «A survey of methods for explaining black box models». In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[15] Lynn Vonder Haar, Timothy Elvira e Omar Ochoa. «An analysis of explainability methods for convolutional neural networks». In: *Engineering Applications of Artificial Intelligence* 117 (2023), p. 105606.

[16] Iraklis A Klampanos et al. «ANNETT-O: an ontology for describing artificial neural network evaluation, topology and training». In: *International Journal of Metadata, Semantics and Ontologies* 13.3 (2019), pp. 179–190.

[17] Aleksandr Kuznetsov, Andrea Maranesi e Alessandro Muscatello. «Deep Learning Based Face Liveliness Detection». In: *2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T)*. IEEE. 2022, pp. 427–432.

[18] Shane T Mueller et al. «Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI». In: *arXiv preprint arXiv:1902.01876* (2019).

[19] Mark A. Musen. «The protégé project: a look back and a look forward». In: *AI Matters* 1.4 (2015), pp. 4–12. DOI: 10.1145/2757001.2757003. URL: https://doi.org/10.1145/2757001.2757003.

[20] Meike Nauta et al. «From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai». In: *ACM Computing Surveys* 55.13s (2023), pp. 1–42.

[21] Meike Nauta et al. *Overview of Methods on Explainable AI*. 2023. URL: https://utwente-dmb.github.io/xai-papers/#/.

[22] Wojciech Samek e Klaus-Robert Müller. ≪Towards explainable artificial intelligence≫. In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 5–22.

[23] Timo Speith. ≪A review of taxonomies of explainable artificial intelligence (XAI) methods≫. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 2239–2250.

[24] *The Digital Europe Programme*. European Commission, 2021. URL: https://digital-strategy.ec.europa.eu/en/activities/digital-programme.

[25] Cornell University. *arXiv Dataset*. Version 143. URL: https://www.kaggle.com/datasets/Cornell-University/arxiv.