European Commission

L’IA “spiegabile” nell'istruzione: promuovere la sorveglianza umana e la responsabilità condivisa

European Digital Education Hub ,Gruppo di lavoro sull'IA spiegabile nell'istruzione

Istruzione e formeazione

EUROPEAN
DIGITAL
EDUCATION
HUB

EUROPEAN
DIGITAL
EDUCATION
HUB

Autori

Francisco Bellas
Jeroen Ooge
Lezel Roddeck
Hasan Abu Rashheed
Marjana Prifti Skenduli
Florent Masdoum
Nurkhamimi bin Zainuddin
Jessica Niewint Gori
Eamon Costello
Lidija Kralj
Deepti Teresa Dcosta Dora Katsamori
Darren Neethling
Sarah ter Maat
Roy Saurabh
Rena Alasgarova
Elena Radaelli
Ana Stamatescu
Arjana Blazic
Graham Attwell
Giedrė Tamoliūnė
Teodora Tziampazi
Moritz Kreinsen
António José Alves Lopes Jose Viñas Diéguez
Barbara Loranc
Cristina Obae

L'European Digital Education Hub (EDEH) è una comunità online che mette in contatto i professionisti di tutte le aree dell'istruzione e della formazione, allo scopo di contribuire al miglioramento dell'istruzione digitale in Europa. Per raggiungere tale obiettivo, EDEH non offre soltanto uno spazio per la discussione e lo scambio di idee, ma inoltre realizza una serie di attività e di iniziative. Tra queste ultime figura la formazione di "squadre", ovvero gruppi di lavoro online attraverso i quali i membri della comunità possono collaborare su un tema specifico legato all'istruzione digitale. Il presente rapporto è il risultato del lavoro della squadra EDEH impegnata sul tema dell'IA spiegabile nell'istruzione.

EUROPEAN
DIGITAL
EDUCATION
HUB

Traduzione libera dall'originale in lingua inglese a cura di Jessica Niewint e Francesca Pestellini (INDIRE*)*

Indice

1. Introduzione all'intelligenza artificiale spiegabile e alle sue implicazioni nell'istruzione

1.1 Contesto

L'intelligenza artificiale spiegabile (XAI) è un sottocampo dell'intelligenza artificiale (IA) che mira a fornire spiegazioni sui motivi per cui un sistema basato sull'IA prende una decisione o fornisce un output (TechDispatch, 2023). La ricerca di spiegazioni comprensibili del funzionamento dei sistemi IA non è nuova, ma in passato si è trattato soprattutto di un'esigenza tecnica degli sviluppatori, che cercavano affidabilità nei risultati ottenuti dai loro sistemi, affinché potessero essere accettati dagli utenti finali in ambiti specifici (Ali et al, 2023). La grande evoluzione della tecnologia IA negli ultimi anni ha trasformato questi sistemi in strumenti digitali di uso comune, facendo emergere nuovi aspetti da prendere in considerazione.

In termini di principi etici nel contesto dell'IA, gli "Orientamenti etici per un'IA affidabile" , pubblicati nel 2019 dal Gruppo di esperti ad alto livello sull'intelligenza artificiale della Commissione europea, hanno stabilito sette requisiti fondamentali per un'IA affidabile: (1) intervento e sorveglianza umani; (2) robustezza tecnica e sicurezza, (3) riservatezza e governance dei dati, (4) trasparenza, (5) diversità, non discriminazione ed equità, (6) benessere sociale e ambientale e (7) accountability. In tale documento di carattere generale si legge:

> *(99) Affinché un sistema sia affidabile, occorre essere in grado di capire perché si è comportato in un certo modo e perché ha fornito una data interpretazione. Un intero campo di ricerca, Explainable IA (XAI) cerca di affrontare questo problema per comprendere meglio i meccanismi alla base del sistema e trovare soluzioni*

Pertanto, la XAI è un campo fondamentale per garantire l'affidabilità dell'intelligenza artificiale e, nel corso della presente trattazione emergerà chiaramente come la XAI fornisca un supporto pratico per la maggior parte dei requisiti etici precedentemente indicati.

Passando alla normativa sull'IA, occorre notare che essa non stabilisce esplicitamente che l'IA debba essere esplicabile, ma piuttosto fa riferimento alla continua sorveglianza dell'essere umano, alla governance dei dati, alla cybersicurezza e alla trasparenza, oltre che al diritto di spiegazione dei singoli processi decisionali.

L'importanza della XAI è aumentata enormemente e in breve tempo sono stati pubblicati nuovi articoli di ricerca e di discussione che ne studiano l'impatto in diversi ambiti (Longo et al, 2024). Tale questione appare rilevante per i policy maker, gli sviluppatori di intelligenza artificiale e gli esperti del settore tecnologico; tuttavia, la maggior parte degli utenti finali non ne è ancora pienamente consapevole e manifesta la necessità di risposte chiare e accessibili a interrogativi fondamentali quali: "Perché sono necessarie queste spiegazioni?" e "Per quale motivo questa tecnologia avanzata risulta non sempre affidabile?" Il rapido progresso dell'informatica negli ultimi 30 anni ci ha portati dai primi personal computer con applicazioni per il calcolo, la creazione di documenti e l'archiviazione di informazioni, ad avere oggi diversi tipi di dispositivi con comunicazione globale permanente e una moltitudine di strumenti in grado di svolgere compiti sempre più sofisticati.

Ma noi, in quanto utenti di tali strumenti, non chiediamo spiegazioni. Ad esempio, nessuno si chiede perché un software di fotoritocco rimuove lo sfondo in un modo specifico o perché viene suggerita una particolare emoji quando si digita una parola in una chat. Per capire perché la tecnologia IA è diversa e richiede spiegazioni possiamo analizzare una definizione generale di intelligenza artificiale, inclusa nell'articolo 3 dell'AI Act :

> *«Sistema di IA»: un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali;*

In questo contesto emergono due concetti chiave. In primo luogo, i sistemi di intelligenza artificiale possono operare in modo autonomo e adattivo, affrontando compiti precedentemente svolti dagli esseri umani, con il rischio di compromettere l'agenzia umana. In secondo luogo, i sistemi di intelligenza artificiale sono in grado di combinare molteplici input e generare output complessi, difficilmente realizzabili e/o comprensibili per gli esseri umani. Tuttavia, tali sistemi risultano meno capaci di ragionare come un essere umano e di mostrare empatia, sensibilità o comprensione delle sfumature culturali. Di conseguenza, se l'IA dovesse prendere decisioni al nostro posto e tali decisioni risultassero poco comprensibili, il rischio insito sarebbe evidente, così come la necessità di ottenere spiegazioni chiare che garantiscano che l'IA stia assistendo e non decidendo autonomamente.

Inoltre, è importante essere consapevoli del fatto che i risultati prodotti dai sistemi di IA potrebbero essere imprecisi (UNU, 2024). La complessità dei problemi affrontati e il funzionamento interno di alcune tecniche utilizzate in questo campo implicano che non possiamo fidarci completamente delle risposte fornite. Per gli utenti comuni della tecnologia digitale si tratta di uno scenario nuovo, poiché le applicazioni tradizionali svolgono o meno il loro compito, ma non esiste una probabilità di successo. Pertanto, i sistemi di IA devono includere spiegazioni adeguate sull'accuratezza dei risultati raggiunti, in modo da rafforzarne l'affidabilità.

L'attuale realtà è che la maggior parte degli utenti non è consapevole della necessità di sistemi di intelligenza artificiale spiegabili, né del fatto che tali funzionalità saranno sempre più integrate nei futuri strumenti di IA per ragioni etiche e giuridiche. Il pubblico generalista, infatti, non sa quali domande porre ai sistemi di IA, né possiede una formazione adeguata per comprenderne correttamente le spiegazioni. È in questo contesto che l'educazione assume un ruolo fondamentale, fornendo alle persone le competenze e le conoscenze necessarie per valutare l'affidabilità dei sistemi di IA, promuovendo il pensiero critico e l'agenzia umana. Ciò è in linea con gli Orientamenti etici per gli educatori sull'uso dell'intelligenza artificiale e dei dati nell'insegnamento e nell'apprendimento (2022), che sottolineano l'importanza di responsabilizzare gli studenti attraverso l'istruzione affinché interagiscano con i sistemi di IA in modo critico, informato e responsabile. Gli Orientamenti saranno rivisti nel 2025. (2022), che sottolineano l'importanza di responsabilizzare gli studenti attraverso l'istruzione affinché interagiscano con i sistemi di IA in modo critico, informato e responsabile. Gli Orientamenti saranno rivisti nel 2025.

Come verrà illustrato nel corso del presente rapporto, l'impatto della XAI nell'istruzione va oltre il mero sviluppo delle competenze. Tuttavia, per inquadrare tale impatto correttamente, è necessario anzitutto analizzare alcune questioni tecniche fondamentali.

1.2 Questioni tecniche

È importante sottolineare che questo è un rapporto a carattere educativo, e non tecnico. Tuttavia, alcuni aspetti tecnici devono essere chiariti per comprendere le specificità di questo ambito. Fornire una spiegazione dell'output generato da un software tradizionale risulta relativamente semplice per gli sviluppatori, in quanto si basa sulla programmazione convenzionale, costituita da una serie di comandi che permettono di analizzare la logica alla base del risultato prodotto. Nel caso dei sistemi di intelligenza artificiale, però, la situazione è più complessa.

In linea generale si distinguono due principali approcci tecnici all'IA: quelli basati sulla conoscenza e quelli basati sui dati (Holmes & Tuomi, 2022). Nel primo caso, la conoscenza e le competenze umane sono rappresentate in una forma che può essere elaborata dai programmi informatici, principalmente attraverso regole logiche e ragionamenti di tipo probabilistico. Questi sistemi erano molto diffusi negli anni Ottanta, ma le difficoltà di scalare a problemi complessi e del mondo reale ne hanno limitato l'impiego a domini controllati. Ottenere spiegazioni dall'IA basata sulla conoscenza è semplice, come nei software standard. Ciò fa capire perché nel campo dell'IA applicata all'istruzione l'approccio basato sulla conoscenza è stato quello più in uso fino alla comparsa dell'IA generativa (Tuomi, 2018). Ad esempio, diversi sistemi di tutoraggio intelligente (ITS), in cui viene creato autonomamente un percorso di apprendimento personalizzato per lo studente, adottano il ragionamento basato su regole *(rule-based reasoning*) e la logica fuzzy. Queste tecniche consentono di includere dashboard dettagliate per insegnanti e studenti, che forniscono spiegazioni visive e tendenze sui progressi dell'apprendimento, aumentando l'affidabilità e l'utilità dei sistemi (Mousavinasab et al, 2018).(ITS), in cui viene creato autonomamente un percorso di apprendimento personalizzato per lo studente, adottano il ragionamento basato su regole e la logica fuzzy. Queste tecniche consentono di includere dashboard dettagliate per insegnanti e studenti, che forniscono spiegazioni visive e tendenze sui progressi dell'apprendimento, aumentando l'affidabilità e l'utilità dei sistemi (Mousavinasab et al, 2018).la logica fuzzy. . Queste tecniche consentono di includere dashboard dettagliate per insegnanti e studenti, che forniscono spiegazioni visive e tendenze sui progressi dell'apprendimento, aumentando l'affidabilità e l'utilità dei sistemi (Mousavinasab et al, 2018).(ITS), in cui viene creato autonomamente un percorso di apprendimento personalizzato per lo studente, adottano il ragionamento basato su regole e la logica fuzzy. Queste tecniche consentono di includere dashboard dettagliate per insegnanti e studenti, che forniscono spiegazioni visive e tendenze sui progressi dell'apprendimento, aumentando l'affidabilità e l'utilità dei sistemi (Mousavinasab et al, 2018).

D'altra parte, l'intelligenza artificiale basata sui dati (data-driven AI) si fonda sull'idea che la conoscenza possa essere estratta direttamente dai dati relativi a un determinato problema, attraverso l'analisi di pattern e l'inferenza, sfruttando l'elevata capacità computazionale dei computer attuali. All'interno di questo approccio, l'apprendimento automatico (Machine Learning; ML) rappresenta il campo di studio specifico in cui vengono sviluppati algoritmi e modelli statistici che i sistemi informatici possono utilizzare per formulare previsioni o prendere decisioni senza ricorrere a istruzioni esplicite. (Marsland, 2011). Oggi disponiamo di algoritmi molto affidabili che adattano i modelli per "apprendere" i pattern nascosti nei dati. Tale processo di apprendimento porta a una serie di parametri numerici che caratterizzano il modello e ne definiscono la risposta. Ottenere spiegazioni da un insieme di numeri non è così semplice come ottenerle da un insieme di regole logiche o comandi scritti in un linguaggio standard, come nel caso dell'IA basata sulla conoscenza. Inoltre, maggiore è il numero di parametri numerici, maggiore è la complessità del modello e, di conseguenza, maggiore è la complessità di ottenere spiegazioni adeguate da esso.

I lettori possono trovare dettagli sulle tecniche, sugli algoritmi e sui modelli specifici associati a ciascuno di questi due approcci in opere di riferimento classiche, come Poole&Mackworth (2023) o Russell&Norvig (2020). Tuttavia, senza la necessità di una comprensione approfondita di tali contenuti, è possibile presentare un semplice esempio che chiarisca le differenze discusse in termini di XAI:

Immaginiamo che un meccanico utilizzi uno strumento di assistenza AI per diagnosticare un problema su un'auto. Come funzionerebbe questo strumento se fosse basato su un'IA basata sulla conoscenza o su un'IA basata sui dati? Per il primo approccio viene utilizzata una tecnica chiamata ragionamento (CBR) basato su casi , mentre per il secondo viene applicato un modello di rete neurale artificiale (ANN). La tabella seguente descrive gli elementi principali del processo decisionale:

	Ragionamento basato su casi (CBR) per l'IA basata sulla conoscenza	**Rete neurale artificiale (ANN) per l'IA basata sui dati**
Scenario	Diagnosticare un problema dell'auto sulla base di un database di casi precedenti.	Diagnosi di un problema all'auto utilizzando caratteristiche di input e previsioni ANN.
Sintomo segnalato	Rumore simile a un clic durante la sterzata.	Rumore simile a un clic durante la sterzata.
Dati di input	Descrizione del rumore e del contesto (ad es. rumore simile a un clic, si verifica durante la svolta).	Codifica numerica delle caratteristiche: • *Tipo di rumore:* clic, cigolio, tonfo (codificato numericamente come 1, 2, 3). • *Azione dell'auto:* svolta, accelerazione, urto contro dossi (codificato numericamente come 1, 2, 3). • *Età dell'auto:* valore numerico (ad es. 5 anni).
Processo	Associa l'input a casi precedenti e applica le regole: • +Regola 1: rumore simile a un clic durante la sterzata → problema al giunto. • Regola 2: rumore stridulo + accelerazione→ problema alla cinghia di trasmissione. • Regola 3: rumore sordo+ urto contro dossi→ problema alle sospensioni.	Elabora gli input attraverso connessioni ponderate e attivazioni: • *Strato di input* (3 nodi): tipo di suono, azione dell'auto, età dell'auto. • *Strato nascosto* (4 nodi): calcola le attivazioni in base ai pesi e alle distorsioni. • *Strato di output* (3 nodi): fornisce una previsione. Probabilità di guasto su giunto, cinghia di trasmissione, sospensioni.
Esempio di corrispondenza	Input: "rumore simile a un clic durante la svolta" corrisponde alla regola 1. Decisione: problema al giunto diagnosticato sulla base di casi precedenti.	Input: "rumore metallico" (1), "svolta" (1), "età dell'auto: 5". Ponderazione tra input e stato nascosto: • Tipo di suono→ nodo nascosto 1: **0,8.** • Azione dell'auto→ nodo nascosto 2: **-0,3.** • Età dell'auto→ nodo nascosto 3: **0,5.** Ponderazione tra strato nascosto e strato di output: • Nodo nascosto 1→ giunto CV: **0,6.** • Nodo nascosto 2→ trasmissione a cinghia: **0,2.** • Nodo nascosto 3→ sospensione: **0,4**. Le somme ponderate e le attivazioni portano a una *probabilità di output*: • Giunto: 85%, cinghia di trasmissione: 10%, sospensione: 5%.
Spiegazione	**Facile da spiegare:** "In un caso precedente con gli stessi sintomi, il problema era il giunto. **Soluzione: sostituire il giunto".**	Difficile da spiegare: dipende dall'interazione tra le ponderazioni e le attivazioni (ad esempio, ponderazione del tipo di suono: 0,85 al nodo del giunto). **Soluzione: probabilità dell'85% di dover sostituire il giunto.**

Tabella 1: Differenze principali tra CBR e ANN nell'esempio di uno strumento di diagnosi automobilistica basato sull'intelligenza artificiale.

Questo esempio mette chiaramente a confronto i due approcci in termini di XAI. Va sottolineato che non tutti i modelli basati sui dati hanno lo stesso livello di opacità, poiché tra le caratteristiche da considerare non vi è soltanto il numero di parametri numerici, ma anche la struttura stessa del modello (Dwivedi et al, 2023).

Tuttavia, i recenti progressi delle tecniche di *machine learning* e il loro successo nell'affrontare problemi del mondo reale sono stati così rilevanti da mettere in ombra l'intelligenza artificiale basata sulla conoscenza. Quando però si tratta di spiegabilità, la situazione rimane ancora incerta, e alcuni settori critici, come l'istruzione e la sanità, continuano a dare priorità a una spiegabilità adeguata, anche a costo di una minore performance del modello (Loh et al. 2022; Khosravi et al, 2022). Di conseguenza, molti ricercatori e sviluppatori di ML stanno lavorando intensamente su tecniche computazionali che consentono di ottenere l'"explainability" (capacità di spiegare) da modelli ML complessi (Bennetot et al, 2024). Si tratta di una ricerca in corso che produrrà alcuni miglioramenti nel prossimo futuro; quindi, dobbiamo essere cauti prima di scartare modelli complessi basati sui dati alla luce della XAI.Khosravi et al, 2022). Di conseguenza, molti ricercatori e sviluppatori nel campo del *machine learning* stanno lavorando intensamente allo sviluppo di tecniche computazionali che permettano di ottenere spiegabilità anche da modelli ML complessi. (Bennetot et al, 2024). La ricerca in questo campo è ancora in corso, e ulteriori miglioramenti sono attesi nel prossimo futuro. Pertanto, è opportuno esercitare cautela nel rifiutare modelli complessi basati sui dati unicamente alla luce delle attuali limitazioni della XAI.

1.3 Definizioni di base

È necessario definire alcuni concetti fondamentali dell'IA spiegabile (XAI), che saranno utilizzati nel corso del presente rapporto. Non è obiettivo di questo lavoro proporre definizioni originali, poiché si tratta di una questione ancora aperta nel dibattito scientifico; ci si baserà pertanto su definizioni già presenti in letteratura, selezionate in base alla loro adeguatezza rispetto al pubblico di riferimento.

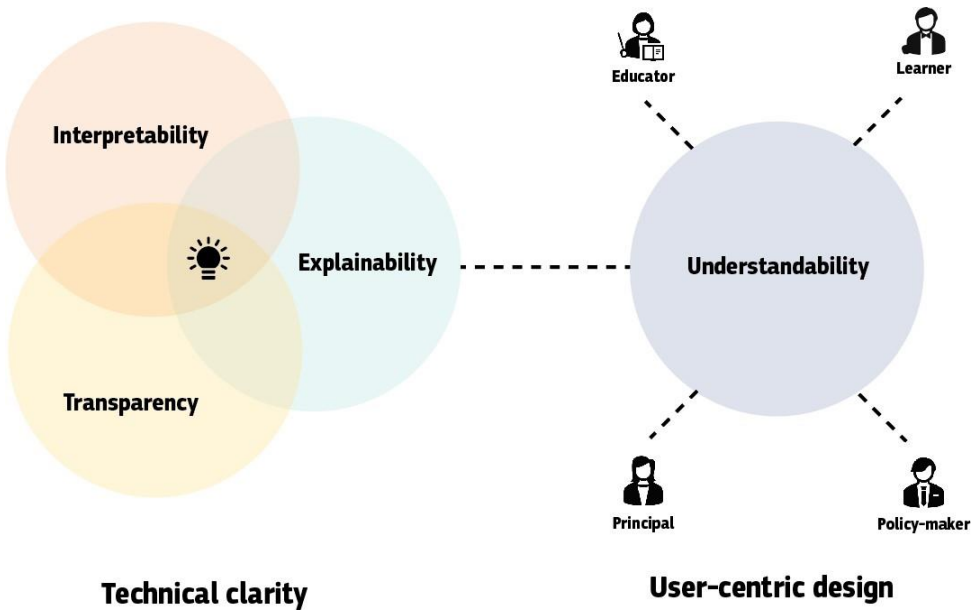

Figura 1: I 4 concetti fondamentali della XAI, organizzati nelle dimensioni "Tecnica" e "Umana"
Fonte: elaborazione degli autori.

Si tratta dei concetti di (1) trasparenza, (2) interpretabilità, (3) spiegabilità e (4) comprensibilità, che verranno definiti nel dettaglio nelle sezioni successive. Anzitutto è importante chiarire che i primi due concetti appartengono alla dimensione tecnica dell'IA, mentre gli ultimi due appartengono alla dimensione umana, come illustrato nella figura 1. A sostegno di quest'ultima, l'obiettivo principale della XAI è che lo sviluppatore includa prima i primi due concetti nel sistema di IA.

Il presente lavoro si allinea alla posizione sostenuta in letteratura (Chaudhry et al, 2022), che considera la trasparenza come il concetto etico fondamentale dell'IA, quale punto di collegamento con gli altri concetti della sicurezza, responsabilità o equità. Per stabilire cosa si intende e per "trasparenza" nell'ambito dell'IA vengono utilizzate due definizioni equivalenti:

Trasparenza

"La trasparenza nell'IA si riferisce a un processo mediante il quale tutte le informazioni, le decisioni, i processi decisionali e le ipotesi vengono rese disponibili per essere condivise con le parti interessate, e tale condivisione contribuisce a migliorarne la comprensione. (Chaudhry et al, 2022)

Ai sensi del considerando (27) dell'AI Act dell'UE, «trasparenza» significa che i sistemi di IA sono sviluppati e utilizzati in modo da garantire un'adeguata tracciabilità e spiegabilità, rendendo gli esseri umani consapevoli del fatto che stanno comunicando o interagendo con un sistema di IA, informando debitamente gli utilizzatori delle capacità e dei limiti di tale sistema e le persone interessate dei propri diritti"

Pertanto, *la trasparenza dipende principalmente*[1] dallo sviluppatore, che deve realizzare il sistema di IA in modo tale da renderlo interpretabile e comprensibile da parte dell'utente. Non si tratta di una caratteristica esclusiva dell'IA, in quanto tale tipo di raccomandazione si ritrova in ambito più generale anche nelle pratiche di scienza aperta promosse dall'UNESCO, che incoraggiano i ricercatori e gli sviluppatori a condividere i dettagli dei loro studi e delle loro scoperte per promuovere l'equità e l'inclusione nell'IA.

Cinque aspetti chiave devono essere contemplati dallo sviluppatore in termini di trasparenza:

1. **Dati:** fornire informazioni sui set di dati utilizzati per addestrare i modelli di IA, comprese le loro fonti, la qualità e le eventuali fasi di pre-elaborazione. Ciò contribuisce a valutare potenziali bias (distorsioni) e la rappresentatività dei dati.
2. **Modello:** offrire approfondimenti sull'architettura, gli algoritmi e i processi decisionali del modello di IA. Ciò consente alle parti interessate di comprendere come gli input vengono trasformati in output, contribuendo alla fiducia e alla responsabilità.

[1] Come si spiega nel prossimo capitolo, l'AI Act conferisce agli individui (utenti finali) il diritto di ottenere dal responsabile dell'implementazione spiegazioni chiare e significative su come il sistema di IA è stato coinvolto nel processo decisionale. Ciò potrebbe essere considerato un livello di trasparenza non tecnico. In ambito educativo, parleremmo della trasparenza degli educatori, in relazione alla loro capacità di spiegare agli studenti, ai genitori o ai colleghi perché stanno utilizzando un determinato strumento di IA (che si ricollega al punto 5 sopra). Si tratta di una considerazione etica che rientra nel concetto di "giustificazione delle scelte" ed è molto rilevante ai fini dell'utilizzo dell'IA nell'istruzione.

3. **Processo:** documentare le procedure di sviluppo e implementazione dei sistemi di IA, comprese le scelte di progettazione, i protocolli di test e gli aggiornamenti. Ciò garantisce che il ciclo di vita dell'IA sia aperto al controllo e in linea con gli standard etici.

4. **Risultato:** comunicare chiaramente i risultati prodotti dai sistemi di IA, insieme ai loro livelli di affidabilità e alle potenziali limitazioni. Ciò aiuta gli utenti a interpretare correttamente i risultati e a prendere decisioni informate.

5. **Scopo:** chiarire l'uso previsto e l'ambito di applicazione del sistema di IA, compresi i suoi obiettivi e il contesto in cui opera. Ciò garantisce che le parti interessate comprendano il ruolo e i limiti dell'IA.

Partendo dal presupposto che la trasparenza sia garantita nella forma poc'anzi descritta e tenendo conto delle questioni tecniche illustrate nella sezione precedente, emerge una dimensione fondamentale della XAI: *l'interpretabilità*.

Interpretabilità

"L'interpretabilità consente **agli sviluppatori** di approfondire il processo decisionale del modello, aumentando la loro fiducia nella comprensione di come il modello ottiene i suoi risultati". (Ali et al, 2023)

"L'interpretabilità si riferisce alla facilità con cui gli esseri umani possono comprendere **il funzionamento di** un **modello** o il modo in cui esso prende decisioni". (Ooge, 2023)

In tal senso, i sistemi di IA possono presentare due livelli principali di interpretabilità (Ooge, 2023):

1. **Modelli intrinsecamente interpretabili:** sono modelli abbastanza semplici da poter essere compresi direttamente dagli esseri umani, come quelli utilizzati dall'IA basata sulla conoscenza (come CBR o regole logiche) e alcuni modelli semplici utilizzati dall'IA basata sui dati (come alberi decisionali o modelli bayesiani). Tali modelli forniscono trasparenza mostrando la logica alla base delle loro decisioni, che è facile da interpretare, e sono denominati modelli *white-box* nel campo della XAI (Ali et al, 2023).

2. **Modelli complessi (opachi):** questi modelli, come le reti neurali o i metodi *ensemble*, sono molto accurati ma difficili da interpretare a causa della loro complessità. Il loro funzionamento interno si basa su grandi insiemi di parametri numerici che vengono regolati utilizzando algoritmi complessi che richiedono lunghi periodi di tempo e molti calcoli. Nel campo della XAI (Ali et al, 2023) vengono solitamente definiti modelli *black-box*.

Nel caso dei modelli opachi, come già evidenziato in precedenza, il campo della XAI è particolarmente attivo nello sviluppo di tecniche di spiegabilità *post hoc*. Ciò significa che l'interpretabilità può essere introdotta successivamente alla fase di addestramento, mediante strumenti quali visualizzazioni, analisi dell'importanza delle caratteristiche (*feature importance*) o con metodi di approssimazione volti a chiarire il comportamento del modello o a giustificare specifiche predizioni (Ooge, 2023). Alcune di queste tecniche saranno spiegate con maggiori dettagli nel prossimo capitolo.

In generale, esiste un compromesso tra le prestazioni del modello e la sua interpretabilità: i modelli più semplici sono più interpretabili ma meno accurati nello svolgimento di compiti complessi. Per una spiegazione più approfondita di questo argomento, si veda la figura 4 in Ali et al (2023).

Spiegabilità

"La spiegabilità fornisce **all'utente finale** informazioni dettagliate sulle decisioni prese dal sistema di IA, al fine di creare fiducia nel fatto che l'IA stia prendendo decisioni corrette e imparziali basate sui fatti". (Ali et al, 2023)

"La spiegabilità nell'IA si concentra sul fornire spiegazioni chiare e coerenti per previsioni o decisioni specifiche del modello. Mira a rispondere a domande come "Perché il sistema di IA ha fatto questa particolare previsione?" offrendo giustificazioni o ragioni **comprensibili all'uomo** per un risultato specifico". (TechDispatch, 2023)

Di conseguenza, il concetto di *interpretabilità* si riferisce alla trasparenza intrinseca dei sistemi di intelligenza artificiale, ovvero alla loro progettazione in modo da non risultare opachi, mentre la *spiegabilità* riguarda la capacità del sistema di giustificare il proprio comportamento agli utenti finali (Hamon et al, 2022; Panigutti et al, 2023). In tale prospettiva, l'interpretabilità è una caratteristica passiva: qualsiasi modello di IA possiede un certo grado intrinseco di interpretabilità (Barredo Arrieta et al. 2020). La spiegabilità, al contrario, è una caratteristica attiva: *un modello di IA è considerato spiegabile quando adotta meccanismi volti a chiarire o rendere comprensibili le proprie funzioni interne, facilitando così la comprensione da parte degli esseri umani (*Barredo Arrieta et al, 2020).

Trasparenza, interpretabilità e spiegabilità rappresentano tre concetti fondamentali della XAI che ricadono principalmente nell'ambito di competenza degli sviluppatori. I primi due si basano su caratteristiche tecniche intrinseche all'approccio adottato, mentre la terza è orientata verso l'utente finale. Al fine di accrescere l'affidabilità percepita del sistema di IA, è essenziale che lo sviluppatore persegua il massimo livello possibile di trasparenza, tenendo tuttavia in considerazione il necessario compromesso tra prestazioni e interpretabilità, poiché il sistema deve mantenere una funzionalità concreta e significativa. Inoltre, lo sviluppatore deve tener conto delle specificità e dei bisogni degli utenti finali, dal momento che la spiegabilità deve essere calibrata sulle loro capacità di comprensione. Questo ultimo aspetto introduce una quarta dimensione della XAI: la comprensibilità.

Comprensibilità

"Il grado in cui le informazioni fornite possono avere senso per le conoscenze **specifiche del pubblico** di destinazione" (Saeed & Omlin, 2023)

"Il grado di **comprensibilità umana** di una decisione presa da un sistema di IA" (TechDispatch, 2023)

Tale caratteristica[2] si riferisce alla misura in cui l'utente finale riesce a comprendere un'esplicazione a lui destinata, costituendo pertanto una misura dell'effettiva utilità della XAI. La comprensibilità si configura, dunque, come una dimensione centrata sull'essere umano, che trascende le caratteristiche puramente tecniche dei modelli di intelligenza artificiale. Si evidenzia la necessità che le spiegazioni dell'IA siano in linea con le esigenze cognitive e contestuali dell'uomo, garantendo che le persone possano comprendere il comportamento del sistema ed i risultati ottenuti (Ooge, 2023)[3],

In sintesi, la XAI si articola attorno a due prospettive principali di sviluppo: quella *tecnica* e quella *umana*. Per chiarire con un caso semplice come i quattro concetti fondamentali della XAI (trasparenza, interpretabilità, spiegabilità e comprensibilità) influenzano lo sviluppo di un sistema di IA, la tabella 2 riprende l'esempio precedente di uno strumento di diagnosi automobilistica basato sull'IA, illustrando come tali concetti possano essere integrati dallo sviluppatore per favorire l'affidabilità del sistema:

[2] La comprensibilità è talvolta definita come equivalente all'interpretabilità nell'ambito della letteratura tecnica sulla XAI (Saeed & Omlin, 2023; Chaudhry et al, 2022). Questi autori considerano lo sviluppatore come l'utente finale, quindi l'interpretabilità dei modelli è correlata alla loro comprensibilità. Ma qui partiamo dal presupposto che la comprensibilità dipende dalla spiegazione adattata al tipo di utente finale, mentre l'interpretabilità è più generale e dipende dal tipo di modello di IA e dalla trasparenza fornita.

[3] Per quanto riguarda il concetto di trasparenza non tecnica introdotto sopra, la comprensibilità è molto importante, poiché consente agli operatori di comprendere lo scopo del sistema di IA e, di conseguenza, di essere trasparenti con l'utente finale. Nell'ambito dell'istruzione, ciò significa che una spiegazione adeguata per gli educatori li aiuta a essere trasparenti con gli studenti, i genitori o i colleghi

	Prospettiva tecnica	Prospettiva umana
Trasparenza	Utilizzare set di dati chiari e ben documentati, come le cronologie delle riparazioni e i dati dei sensori di vari modelli di auto.	Informare i meccanici e i proprietari delle auto sulle fonti dei dati utilizzati per la diagnosi (ad esempio, "basato su 10.000 riparazioni di auto").
	Condividere i tipi di problemi che lo strumento è in grado di diagnosticare (ad esempio, guasti al motore, stato della batteria) e i suoi limiti.	Pubblicare un manuale che spieghi l'ambito di applicazione dello strumento e assicurarsi che gli utenti comprendano che si tratta di uno strumento di supporto e non di una fonte autorevole di riferimento.
Interpretabilità	Scegliere un modello interpretabile per questioni più semplici (ad esempio, alberi decisionali per lo stato della batteria).	Fornire ai meccanici strumenti che mostrano percorsi decisionali chiari, come "Bassa tensione nella cella 3, probabilità del 75% che la batteria sia difettosa".
	Per diagnosi più complesse (ad esempio, accensioni irregolari del motore), utilizzare strumenti di valutazione dell'importanza delle caratteristiche per evidenziare i fattori chiave.	Formare i meccanici su come interpretare e verificare i livelli di affidabilità del modello di IA e i dati di sensibilità con controlli fisici o ulteriori test.
Spiegabilità	Includere strumenti di spiegazione che mostrano perché il sistema suggerisce problemi specifici (ad esempio, "in base alle fluttuazioni del numero di giri del motore").	Fornire supporti visivi, come diagrammi annotati, che spieghino le parti dell'auto interessate (ad esempio, "L'IA ha rilevato una perdita nel sistema di iniezione del carburante, con un'elevata affidabilità").
	Utilizzare spiegazioni controfattuali: "Se la tensione della candela fosse più alta, questo problema potrebbe non verificarsi".	Assicurare che le spiegazioni siano facili da capire per i proprietari delle auto, concentrandosi sulle azioni da intraprendere (ad esempio, "Sostituisci la candela, l'accuratezza di questa previsione è elevata").
Comprensibilità	Semplificare il linguaggio nell'interfaccia (ad esempio, "Rilevato un guasto nel sistema di scarico" invece di "Problema di ricircolo dei gas di scarico").	Fornire un'app o una dashboard chiara e intuitiva che consenta al proprietario dell'auto di visualizzare le diagnosi con i livelli di gravità (ad esempio, "critico, riparazione necessaria, alta affidabilità").
	Utilizzare elementi visivi (ad esempio, diagrammi di sistema) per evidenziare le aree problematiche.	Coinvolgere meccanici e proprietari di auto durante i test per garantire che le spiegazioni siano utili e attuabili.
Integrazione	Monitorare continuamente le prestazioni dello strumento con il feedback dei meccanici. Aggiornare i modelli secondo necessità per ridurre le diagnosi errate.	Fornire formazione continua ai meccanici e assistenza clienti ai proprietari di auto. Raccogliere regolarmente feedback per migliorare la chiarezza e la funzionalità.

Tabella 2: Prospettiva tecnica e umana per l'esempio di strumento di diagnosi automobilistica basato sull'intelligenza artificiale.

1.4 Caratteristiche principali delle spiegazioni nei sistemi di IA

È importante convergere su una serie di caratteristiche fondamentali che le spiegazioni dovrebbero idealmente presentare per soddisfare i requisiti di responsabilità etica e di effettiva utilità per l'utente finale.

Gli sviluppatori dovrebbero tenere conto di tali dimensioni nella fase di progettazione dei sistemi di intelligenza artificiale.

Un “insieme minimo” di queste caratteristiche può essere sintetizzato come segue:

Categoria	Sotto-caratteristica	Descrizione	Esempio (istruzione)
Chiarezza (per favorire la comprensibilità)	**Linguaggio semplice**	La spiegazione dovrebbe evitare termini tecnici	*"Il nostro sistema ha esaminato i tuoi recenti punteggi nei quiz e ha notato che hai avuto difficoltà con le equazioni algebriche. Potrebbe esserti utile ripassare quei concetti specifici".*
	Livelli multipli di dettaglio	Utenti diversi potrebbero aver bisogno di maggiori o minori dettagli, quindi le spiegazioni dovrebbero fornire inizialmente informazioni di base, con la possibilità di approfondire, all'occorrenza, gli aspetti tecnici o relativi ai dati	*Livello base: "Abbiamo utilizzato i risultati del quiz per identificare le aree di miglioramento". Livello dettagliato: "Abbiamo combinato i punteggi di più quiz e ponderato ogni domanda in base alla difficoltà per determinare che la scomposizione algebrica è la tua competenza più debole".*
Pertinenza (in relazione al contesto)	**Linguaggio semplice**	Le spiegazioni devono essere pertinenti al contesto dell'applicazione	*"Poiché il tuo saggio presentava ripetuti errori grammaticali, il sistema suggerisce di esercitarsi ulteriormente sulla struttura delle frasi, che è fondamentale per questo corso di scrittura inglese".*
	Livelli multipli di dettaglio	La spiegazione dovrebbe aiutare l'utente a compiere i passi successivi o a prendere decisioni pratiche	*"Sulla base dei risultati del quiz, il sistema consiglia di rivedere il capitolo 4 del libro di testo e di completare gli esercizi pratici entro venerdì."*
Specificità (relativa alla tecnologia AI)	**Processo decisionale o logica del modello**	L'utente deve sapere che il sistema utilizza determinati dati inseriti e un modello di *machine learning* (ML) o un modello basato su regole per generare raccomandazioni o decisioni.	*"Abbiamo addestrato un modello ML sui punteggi dei quiz e sui voti finali degli studenti precedenti. I tuoi dati attuali sulle prestazioni sono stati confrontati con profili di studenti simili per suggerirti aree di studio mirate".*
	Limiti e incertezze	La spiegazione dovrebbe indicare il margine di incertezza del risultato, per esempio includendo livelli di affidabilità o menzionando situazioni in cui i dati potrebbero essere incompleti o distorti.	*"Questa raccomandazione potrebbe non riflettere appieno la tua comprensione se non hai ancora completato tutti i quiz. Il livello di affidabilità della raccomandazione è dell'86%".*
Tracciabilità (per la responsabilità)	**Chi è responsabile**	L'utente deve essere in grado di identificare chi (o quale organizzazione) è responsabile dei risultati del sistema e chi contattare per eventuali chiarimenti.	*"Questo sistema di raccomandazioni è gestito dall'Ufficio.... Per qualsiasi domanda o dubbio, contattare [email]".*
	Verificabilità	Il sistema dovrebbe registrare le fasi del processo decisionale e i dati in modo da permettere a un audit interno o esterno di verificare come sono state raggiunte determinate conclusioni	*"Tutti i dati utilizzati per generare la raccomandazione vengono registrati. Un comitato per l'integrità accademica potrà esaminare il registro per garantire che i suggerimenti siano stati prodotti in modo equo e accurato".*
Coerenza (affidabilità)	**Spiegazioni coerenti**	Le spiegazioni per casi o input simili non dovrebbero variare in modo significativo; dovrebbero infatti seguire la stessa logica o le stesse regole.	*"Gli altri studenti che hanno avuto difficoltà specifiche con la scomposizione dei polinomi hanno ricevuto la stessa raccomandazione sul modulo di studio, garantendo coerenza tra profili simili".*
	Nessun messaggio contraddittorio	Se esistono più "livelli" di spiegazione (di base e dettagliata), questi non devono essere in conflitto tra loro.	*"La panoramica di alto livello rileva che l'algebra è la tua sfida principale e l'analisi dettagliata conferma che la scomposizione dei polinomi è l'area di competenza chiave che necessita di revisione".*

Tabella 3: Caratteristiche principali delle spiegazioni nell'IA.

1.5 Prospettive e livelli di XAI

Alla luce delle considerazioni finora esposte, risulta evidente che il progresso nel campo della XAI richiede il coinvolgimento di una pluralità di attori e portatori di interesse, nonché la promozione della loro collaborazione. Come punto di partenza, si potrebbe adottare l'approccio proposto da Saeed e Omlin (2023), che prende in esame cinque prospettive distinte e le associa a tre principali categorie di stakeholder, ciascuna con un ruolo differente nel contesto della XAI.:

6. **Legislatori e policy maker:** con prospettive normative e sociali. L'approccio dell'UE alla regolamentazione del digitale si basa sui diritti fondamentali delle persone e mira a incoraggiare l'innovazione promuovendo un'IA affidabile e centrata sull'uomo. Si tratta di una strategia equilibrata, che promuove lo sviluppo di un'IA sicura ed etica, con particolare attenzione alle finalità per cui i sistemi sono stati progettati. Pertanto, il ruolo del legislatore e del decisore politico non si limita al controllo dello sviluppo tecnologico, ma si estende, in misura ancora più rilevante, alla definizione della governance e alla sorveglianza sumana sull'intero sistema di IA. Tali attori hanno infatti la facoltà di definire normative e politiche che regolamentano lo sviluppo e l'impiego dell'IA.

7. **Ricercatori, professionisti e sviluppatori:** con prospettive sia industriali (professionali) sia legate allo sviluppo dei modelli — rappresentano la forza tecnica trainante della XAI e i principali responsabili del progresso del settore. È a questo livello che devono essere gestiti gli interessi commerciali, perseguendo un equilibrio tra vincoli normativi e opportunità di innovazione. A tali stakeholder spetta il compito di promuovere la trasparenza, l'interpretabilità, la spiegabilità e la comprensibilità dei sistemi di intelligenza artificiale.

8. **Utenti finali:** con prospettive industriali (professionali) e sociali. Questo gruppo comprende un pubblico eterogeneo, che può applicare i sistemi di IA per il proprio sviluppo professionale o per questioni specifiche. Anche gli utenti finali devono contribuire a promuovere la spiegabilità e la comprensibilità dei sistemi di IA. Come commentato in precedenza, solo quest'ultima dimensione dipende dall'utente finale, ma è la più rilevante, poiché è qui che risiede l'obiettivo finale dell'affidabilità. Le istituzioni appartenenti ai settori dell'istruzione e della formazione rientrano in questa categoria.

Da un punto di vista pratico, anche i primi due tipi di stakeholder sopra indicati sono utenti finali dell'IA e,di conseguenza, sono influenzati dalla spiegabilità e comprensibilità dei sistemi che promuovono e sviluppano. In tal senso, è possibile distinguere tre livelli di XAI, adattati da Ooge, 2023:

Principianti: individui interessati dai sistemi di IA con competenze tecniche minime o nulle in materia. Hanno bisogno di spiegazioni per comprendere meglio i modelli di IA, garantendo il rispetto dei principi di equità, fiducia e riservatezza dei dati.

Esempi: pazienti, richiedenti prestiti, organismi di regolamentazione, personale amministrativo, studenti, insegnanti.

Utenti di livello avanzato: professionisti come data scientist e specialisti di settore che utilizzano l'IA per l'analisi e il processo decisionale ma non dispongono di competenze tecniche approfondite in materia. Hanno bisogno di strumenti avanzati per valutare l'affidabilità dei modelli, ottimizzarli e confrontarli.

Esempi: medici, funzionari di banca, manager, giudici, assistenti sociali, ricercatori nel campo dell'istruzione, responsabili dei sistemi informatici scolastici, insegnanti di informatica.

Esperti di IA: specialisti che creano e implementano modelli di IA o sviluppano tecniche di IA spiegabili. Si concentrano sull'interpretazione e il miglioramento dei loro modelli per garantirne il corretto funzionamento.

Esempi: ricercatori di IA, ingegneri.

1.6 La XAI nell'istruzione

Le sezioni precedenti hanno fornito una panoramica generale sulla XAI. Tuttavia, poiché il presente rapporto è incentrato *sull'istruzione*, è necessario inquadrare la XAI in tale ambito specificp.

In ambito educativo, l'intelligenza artificiale viene vista sempre più spesso come uno strumento promettente, con il potenziale di migliorare le esperienze di apprendimento, affrontare le sfide e personalizzare l'insegnamento per supportare meglio le esigenze diverse e in continua evoluzione degli studenti, sebbene le sue piene capacità e il suo impatto pratico siano ancora in fase embrionale. Per gli studenti, i sistemi di intelligenza artificiale mirano a fornire esperienze di apprendimento personalizzate, analizzando i punti di forza e di debolezza individuali, fornendo contenuti su misura, offrendo feedback in tempo reale e identificando le aree di miglioramento per colmare efficacemente le lacune di conoscenza. Per gli educatori, l'IA ha il potenziale per gestire compiti di routine come la valutazione e la pianificazione delle lezioni, liberando tempo utile da dedicare alle interazioni interpersonali con gli studenti. L'IA può promuovere l'inclusività sostenendo gli studenti con disabilità, gli studenti multilingue e coloro che necessitano di formati di apprendimento alternativi. Per una solida introduzione dell'IA nell'istruzione, si veda la relazione del precedente gruppo EDEH sull'IA, che costituisce la base del presente rapporto.

Massime della spiegabilità dell'IA		
1) Trasparenza		**2) Responsabilità**
La trasparenza nei sistemi di intelligenza artificiale implica la comunicazione chiara delle loro funzionalità e dei processi decisionali. Le organizzazioni sono tenute a dichiarare l'utilizzo dell'IA e a fornire spiegazioni comprensibili per un pubblico eterogeneo, nel rispetto delle normative vigenti, come il GDPR. Tale trasparenza favorisce la fiducia e consente agli stakeholder di comprendere i risultati generati dal sistema.		La responsabilità nella governance dell'IA definisce ruoli e obblighi lungo l'intero ciclo di vita del sistema. Le organizzazioni devono giustificare le scelte progettuali e operative, offrendo agli stakeholder la possibilità di contestare decisioni, ad esempio in conformità con l'AI Act e il GDPR. Questo rafforza la supervisione etica.
3) Considerazione del contesto		**4) Riflessione sugli impatti**
Le spiegazioni devono essere adattate alle esigenze specifiche del pubblico di riferimento, che può includere insegnanti, studenti o decisori politici. Questa adattabilità contestuale garantisce che la comunicazione sui sistemi di IA sia pertinente e significativa, tenendo conto delle competenze e del dominio applicativo degli stakeholder.		I sistemi di IA dovrebbero concentrarsi sull sistemi di IA devono essere orientati al benessere umano e sociale, mediante valutazioni continue dei benefici e dei rischi. Le organizzazioni devono adottare strategie per mitigare i potenziali danni e promuovere l'inclusività, assicurando che l'IA sia allineata a obiettivi etici e concreti.

Figura 2: Massime di spiegabilità dell'IA.
Fonte: lavoro degli autori.

Garantire la spiegabilità dei sistemi di IA è fondamentale per lo sviluppo di sistemi etici e affidabili, in particolare nel settore dell'istruzione, dove le decisioni producono effetti importanti sia sul singolo che sulla società nel suo complesso. Di conseguenza, sono emerse sfide giuridiche, etiche e di governance di una certa complessità, in particolare per quanto riguarda la trasparenza, l'equità e la responsabilità ("TFA"). In quest'ottica, l'Alan Turing Institute ha presentato il quadro "AI Explainability in Practice", che propone quattro principi guida per costruire una solida base volta a garantire che i sistemi di IA siano trasparenti, responsabili e in linea con le esigenze dei vari soggetti interessati (vedi figura 2). Tali principi colmano il divario tra gli obblighi di natura legale, le considerazioni etiche e la messa in pratica, promuovendo la fiducia e l'usabilità in contesti diversi. I principi forniscono inoltre orientamenti chiari su come comunicare efficacemente alle persone le decisioni assistite dall'IA, garantendo c h e le spiegazioni siano significative e in linea con le diverse esigenze degli stakeholder.

Impatto specifico della XAI nel settore dell'istruzione

La XAI influisce sull'istruzione in molti modi. In linea di massima, potremmo considerare due aspetti principali:

Sviluppo delle competenze: tutti gli attori coinvolti nell'istruzione devono possedere competenze adeguate in materia di IA, tra le quali conoscenze, abilità e attitudini (Vuorikari et al, 2022). È fondamentale esercitare il pensiero critico e preservare l'autonomia decisionale nell'utilizzo degli strumenti di intelligenza artificiale nei contesti di insegnamento, apprendimento o gestione amministrativa.

Inoltre, nel caso degli educatori, è necessario sviluppare competenze in materia di XAI per comprendere adeguatamente le caratteristiche di spiegabilità degli strumenti utilizzati in classe.

Sviluppo di strumenti di IA per l'istruzione: gli sviluppatori di strumenti di IA in Europa devono rispettare gli obblighi di trasparenza e spiegabilità stabiliti dall'AI Act. Nel caso dell'istruzione, tali requisiti devono essere discussi e concordati con educatori e pedagogisti, poiché, per quanto riguarda gli studenti, la comprensibilità e l'affidabilità delle spiegazioni possono interferire con il processo di apprendimento. L'apprendimento autoregolato - inteso come processo attivo in cui gli studenti utilizzano le loro capacità cognitive e fisiche per acquisire competenze rilevanti al fine di svolgere compiti specifici -, potrebbe essere compromesso se la XAI non fosse implementata correttamente (Azfaal et al, 2023). Pertanto, le autorità educative dovrebbero tenere conto delle caratteristiche specifiche degli studenti in termini di spiegabilità e comprensibilità, quando scelgono gli strumenti di intelligenza artificiale da introdurre negli ambienti di apprendimento. Si rafforza così la conclusione della sezione precedente, secondo cui nel campo dell'istruzione la XAI richiede una stretta cooperazione tra gli attori coinvolti nella fase di sviluppo e una supervisione attiva da parte delle autorità accademiche nella fase di implementazione.

Più specificamente, l'istruzione presenta esigenze particolari in materia di XAI (Khosravi et al, 2022):

Responsabilità

Gli educatori devono essere responsabili nei confronti degli studenti, dei genitori o dell'amministrazione quando utilizzano sistemi di IA per l'insegnamento, l'analisi dell'apprendimento o l'assegnazione di compiti.

Trasparenza

Gli educatori devono spiegare come vengono prese le decisioni e come vengono utilizzati i sistemi di IA nelle istituzioni scolastiche, in particolare quando trattano dati personali o utilizzano sistemi di IA come definiti dalle normative vigenti.

Metacognizione e agenzia

Le spiegazioni devono consentire agli studenti di assumere un maggiore controllo sul proprio apprendimento promuovendo l'autoriflessione, la pianificazione e il processo decisionale.

Conformità dei dati ai requisiti di legge: I fornitori e gli utilizzatori di sistemi di IA devono rispettare le disposizioni di legge, quali quelle contenute nel GDPR, garantendo la trasparenza in materia di raccolta, utilizzo e conservazione dei dati, con particolare attenzione alla protezione dei minori.

Gestione di dati rumorosi e complessi: dal punto di vista tecnico, i dati didattici relativi al processo di apprendimento dello studente provengono spesso da fonti eterogenee, tra le quali interazioni digitali, valutazioni e osservazioni comportamentali. Tali dati sono frequentemente caratterizzati da rumore e ambiguità, e richiedono pertanto un'analisi accurata e metodologicamente rigorosa per poter essere interpretati in modo affidabile**.**

Fraintendimenti: i sistemi di intelligenza artificiale applicati all'ambito educativo devono essere progettati in modo da prevenire l'introduzione di concetti errati o comportamenti di apprendimento indesiderati, che possono derivare da inesattezze o bias presenti negli output generati. Tale rischio è una conseguenza diretta della complessità dei dati utilizzati, come precedentemente discusso**.**

Progettazione centrata sulla pedagogia: le spiegazioni fornite dai sistemi di IA dovrebbero essere coerenti con gli obiettivi pedagogici e fondate sui principi delle scienze dell'apprendimento, al fine di massimizzare l'efficacia dei risultati educativi**.**

Stakeholder specifici[4]

Utenti finali: nel contesto educativo, insegnanti e studenti hanno bisogno di ricevere spiegazioni semplificate e chiare per comprendere le raccomandazioni generate dai sistemi di intelligenza artificiale e agire di conseguenza in modo efficace. Gli insegnanti devono essere in grado di valutare criticamente gli output dell'IA, mentre gli studenti richiedono un feedback operativo che promuova fiducia, coinvolgimento e consapevolezza del proprio processo di apprendimento. In linea con quanto emerso durante la pandemia da COVID-19, anche i genitori possono essere inclusi tra gli utenti finali: essi traggono beneficio da informazioni trasparenti sulle prestazioni e sui progressi dei figli, al fine di fornire un supporto educativo più efficace.

Autorità educative: i dirigenti scolastici, i responsabili delle politiche educative e le autorità istituzionali necessitano di spiegazioni che integrino prospettive globali e locali, per poter prendere decisioni informate basate su evidenze affidabili e contestualizzate.

Sviluppatori: gli sviluppatori (o i fornitori, compresi alcuni importatori o distributori definiti dalla legge sull'IA, in particolare per i sistemi ad alto rischio) devono garantire che gli strumenti didattici basati sull'IA forniscano spiegazioni pratiche e di facile comprensione, significative per gli utenti non esperti. La collaborazione con gli educatori è essenziale per garantire che i risultati siano chiari, attuabili e in linea con le esigenze didattiche reali.

[4] Nei prossimi capitoli si farà riferimento a queste principali tipologie di soggetti interessati all'istruzione, ma in ogni contesto specifico potrebbe essere necessario adottare una terminologia leggermente diversa (ad esempio, i termini giuridici potrebbero riferirsi agli studenti e agli educatori come utenti finali, o agli amministratori come autorità educative).

Sfide, limiti e opportunità

Il seguente diagramma SWOT fornisce una sintesi complessiva delle principali sfide, limitazioni e opportunità legate all'integrazione della XAI nel contesto educativo, che saranno approfondite nel corso del presente rapporto.

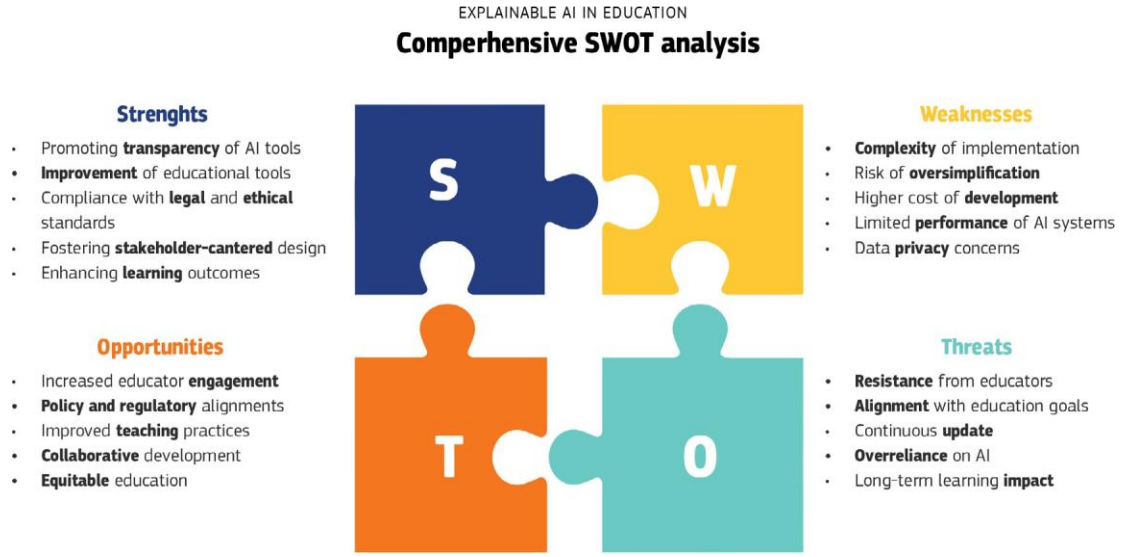

Figura 3: Analisi SWOT completa.
Fonte: elaborazione degli autori.

Una tassonomia della XAI nell'istruzione

La Tabella 4 presenta una tassonomia finalizzata a delineare le dimensioni chiave delle spiegazioni fornite dai sistemi di intelligenza artificiale, con esempi specificamente riferiti ai contesti educativi. La struttura si articola su due livelli di spiegazione: (1) le proprietà del modello o del sistema di IA e (2) le modalità di presentazione delle spiegazioni ai soggetti interessati o agli utenti finali. Ciascuna dimensione ("ambito", "profondità", "alternative" e "flusso") identifica modi specifici in cui le spiegazioni possono essere adattate per soddisfare le esigenze dei diversi soggetti interessati. La tassonomia è stata adattata da Kesari et al (2024) e sarà citata nei prossimi capitoli.

Ambito **Spiegazioni globali o locali**	**Spiegazione globale** Fornisce una comprensione completa del comportamento del modello in un'ampia gamma di scenari Ad esempio, comprendere l'adeguatezza di uno strumento didattico ai requisiti operativi in linea con le politiche istituzionali, come richiesto dal preside della scuola	**Spiegazione locale** Si concentra sulla comprensione del comportamento del modello per un caso specifico o un piccolo insieme di casi Ad esempio, spiegare a un insegnante perché un determinato studente ha ricevuto un voto pari a zero a causa di un sistema di valutazione basato sull'intelligenza artificiale
	Ad esempio, valutare se una piattaforma di attribuzione dei voti basata sull'intelligenza artificiale dimostra un bias costante tra diversi gruppi demografici di studenti, da utilizzare da parte di un responsabile politico	

Profondità: **Livello di selettività nelle spiegazioni**	**Spiegazioni esaurienti** Trasmettere valutazioni complete per revisioni approfondite del sistema	**Spiegazioni selettive** Approfondimenti semplificati per un feedback immediato
Alternative: **Spiegazioni contrastive o non contrastive**	**Spiegazioni contrastive** Evidenziano la differenza tra ciò che è accaduto e ciò che era previsto, concentrandosi sui risultati alternativi Esempio: uno studente chiede perché ha ricevuto un voto inferiore rispetto a un compagno in un compito valutato dall'IA. Ad esempio, un educatore vuole sapere perché un determinato consiglio sul corso è stato fornito a uno studente e non a un altro.	**Spiegazioni non contrastive** Forniscono informazioni sul comportamento del modello senza fare riferimento a un risultato alternativo Esempio: gli educatori che desiderano comprendere i fattori o le caratteristiche che l'IA considera più importanti nell'assegnazione dei voti agli studenti Ad esempio, analizzare i principi generali alla base dei suggerimenti di uno strumento di IA per lo sviluppo del programma di studi
Flusso: **Come vengono trasmesse le spiegazioni (cioè come condizioni o correlazioni)**	**Spiegazioni condizionali** Spiegazioni basate su regole per decisioni mirate (i formati "se-allora" mostrano quando si verificano risultati specifici). Rendono le spiegazioni più chiare e facili da comprendere. Utili per linee guida semplici, chiare e attuabili. L'inconveniente è che semplificano eccessivamente relazioni complesse e non colgono le sottigliezze delle variabili. Ad esempio, un sistema di apprendimento personalizzato raccomanda esercitazioni aggiuntive sulla base di una soglia di punteggio predefinita del 60%: *"Se il punteggio ottenuto dallo studente nel quiz sull'argomento A è inferiore al 60%, ASSEGNA ulteriori esercizi sull'argomento A."*	**Spiegazioni correlazionali** Utili per comprendere come i cambiamenti nei dati di input influenzano l'output del modello. Utili per l'analisi delle tendenze e per ottenere informazioni probabilistiche. Tuttavia, difficili da interpretare. Ad esempio, un sistema di valutazione automatico mostra che i punteggi più alti sono fortemente correlati al tempo dedicato agli esercizi, aiutando gli educatori a comprendere le tendenze sistemiche. Ad esempio, una piattaforma di apprendimento personalizzata consiglia ulteriori materiali di lettura se il rendimento di uno studente nei quiz è in calo. Ciò mostrerebbe al docente una correlazione tra l'aumento del numero di esercizi consigliati e il calo dei punteggi dei quiz.

Tabella 4: Dimensioni chiave della spiegabilità nei sistemi di IA applicati all'istruzione.

La differenza tra la Tabella 3 e la Tabella 4 deve risultare chiara. La Tabella 3 si concentra su ciò che rende una spiegazione eticamente valida e facilmente comprensibile per l'utente. Essa mette in evidenza una serie di qualità operative — simili a una lista di controllo — che ogni spiegazione fornita da un sistema di IA dovrebbe cercare di incorporare. Risulta particolarmente utile per chi implementa soluzioni educative e ha la necessità di valutare rapidamente se una spiegazione soddisfa standard etici e pedagogici fondamentali.
Al contrario, la Tabella 4 offre una visione più ampia delle possibili strategie esplicative, consentendo agli operatori del settore educativo di scegliere l'approccio più adatto alle proprie esigenze (ad esempio, una spiegazione "locale" e immediata per giustificare il voto di un singolo studente, rispetto a una panoramica "globale" utile per definire politiche istituzionali). Entrambe le tabelle possono servire da guida per gli sviluppatori nella progettazione di sistemi di intelligenza artificiale in grado di fornire spiegazioni chiare, pertinenti e adeguate al contesto educativo.

1.7 Finalità e organizzazione del presente rapporto

Il presente rapporto si propone di offrire un contributo alla comunità educativa mediante un'analisi formale e aggiornata delle implicazioni della XAI nel campo dell'istruzione. L'attenzione sarà rivolta a questioni educative di carattere ampio e sistemico, evitando di concentrarsi su strumenti specifici, al fine di adottare un approccio solido e orientato al futuro. Saranno inoltre fornite raccomandazioni pratiche rivolte a tutti gli attori coinvolti nel settore educativo: sviluppatori, autorità educative, insegnanti e studenti.

Il presente rapporto è articolato in tre capitoli principali. Il primo capitolo è dedicato agli aspetti giuridici della XAI nel contesto educativo. Considerando l'attuale evoluzione del quadro normativo, risulta fondamentale analizzare in che modo l'AI Act, il GDPR e altri strumenti di regolamentazione digitale incidano sull'adozione di strumenti trasparenti e spiegabili nell'ambito dell'istruzione. Il capitolo ha l'obiettivo di approfondire la comprensione dell'impatto della XAI nel settore educativo, offrendo ai lettori gli strumenti per cogliere la complessità dell'implementazione dell'intelligenza artificiale in conformità con la normativa europea. I diversi stakeholder — tra cui insegnanti, studenti, sviluppatori e decisori politici — acquisiscono una maggiore consapevolezza rispetto ai propri ruoli e alle rispettive responsabilità in ordine alla spiegabilità dell'IA. Gli sviluppatori, in particolare, possono individuare opportunità per progettare sistemi di IA che coniughino eccellenza tecnica e requisiti di spiegabilità, contribuendo così a rafforzare l'affidabilità percepita delle tecnologie. In sintesi, questo capitolo intende preparare i lettori ad affrontare e gestire in modo proattivo i futuri sviluppi normativi e tecnologici nel panorama dell'educazione digitale.

Il secondo capitolo affronta le problematiche connesse alle diverse prospettive degli utenti. L'importanza della XAI nel contesto educativo viene sottolineata attraverso l'evidenziazione del suo ruolo nel promuovere fiducia, trasparenza e responsabilità tra i vari stakeholder coinvolti. Si tratta di un capitolo di taglio applicativo, incentrato su due principali applicazioni dell'IA in ambito educativo: i sistemi di tutoraggio intelligenti (ITS) e i generatori di piani di lezione basati sull'IA (LPG). Viene analizzata la prospettiva dei diversi utenti nell'utilizzo di tali strumenti, evidenziando esigenze, aspettative e criticità. Da tale analisi emerge chiaramente come il raggiungimento della spiegabilità nei sistemi di IA per l'educazione richieda un approccio collaborativo e centrato sull'essere umano, in cui tutti gli attori — educatori, studenti, sviluppatori, dirigenti e policy maker — siano attivamente coinvolti.

Nell'ultimo capitolo, dedicato *all'alfabetizzazione all'IA e al pensiero critico*, si sottolinea l'importanza di promuovere il pensiero critico come obiettivo educativo fondamentale, utilizzando la XAI per migliorare la comprensione e la trasparenza. Tra i contributi principali vi è la proposta di un insieme di competenze fondamentali per gli educatori, applicabili a tutti i livelli di istruzione, finalizzate a favorire la comprensione, la valutazione e l'implementazione della XAI nei contesti educativi.

Il capitolo introduce, come nuovo aspetto, una serie di competenze che possono essere utili agli insegnanti per comprendere le dimensioni chiave della XAI illustrate nella tabella 4. Esempi pratici illustreranno l'integrazione della XAI nei programmi di studio, dall'istruzione primaria a quella superiore e alla formazione professionale, con attività volte a demistificare i processi di IA e a promuovere un impegno critico.

2. Orientarsi nel rispetto dell'AI Act, del GDPR e delle normative correlate sul digitale

2.1 Contesto

Il regolamento UE sull'intelligenza artificiale (in prosieguo: il «regolamento sull'IA» o l'«AI Act»), il regolamento generale sulla protezione dei dati (GDPR) e le leggi correlate[5] regolano gli spazi digitali in cui interagiscono diversi stakeholder — tra i quali studenti, insegnanti, aziende edtech e autorità educative. I sistemi di intelligenza artificiale operano all'interno di ecosistemi complessi, in cui tali attori richiedono livelli differenziati di spiegazione. Gli studenti, ad esempio, possono necessitare di motivazioni semplici e accessibili rispetto alle decisioni prese dal sistema — come la proposta di un nuovo esercizio allo stesso livello di difficoltà anziché il passaggio a uno superiore —, mentre gli insegnanti richiedono informazioni più specifiche che consentano di allineare le raccomandazioni dell'IA agli obiettivi pedagogici. D'altro canto, gli sviluppatori di tecnologie educative e gli enti regolatori necessitano di spiegazioni dettagliate, incentrate sui processi, per garantire accuratezza tecnica e conformità agli standard etici e normativi. Di conseguenza, una singola forma di spiegazione non può soddisfare tutte queste esigenze: è necessario adottare quadri di spiegabilità, flessibili e multilivello. A ciò si aggiunge una sfida tecnica rilevante, ovvero la traduzione dei meccanismi complessi degli algoritmi — in particolare quelli basati su tecniche avanzate come le reti neurali — in spiegazioni comprensibili anche da parte di utenti non esperti.

Per garantire che le normative vigenti siano effettivamente attuate e rispettate dagli stakeholder, è necessario che i sistemi di intelligenza artificiale risultino comprensibili, utilizzabili e pertinenti rispetto al destinatario cui si rivolgono. Di conseguenza, lo sviluppo di metodi che rendano accessibile la logica sottostante ai sistemi di IA anche a soggetti non esperti — senza però incorrere in semplificazioni eccessive o in distorsioni dei processi interni — rappresenta una sfida significativa. Affrontare tali complessità nei sistemi di IA destinati all'istruzione richiede un approccio equilibrato*: i metodi di spiegazione devono semplificare tecniche complesse in modo tale da favorire la comprensione e la fiducia da parte degli utenti coinvolti nei processi decisionali — come studenti, insegnanti e personale amministrativo —, rispettando al contempo i requisiti giuridici e le esigenze specifiche del settore educativo*. Il presente capitolo è strutturato secondo una progressione logica. Si apre con un'introduzione al contesto educativo e ai bisogni concreti legati all'utilizzo dell'IA nelle istituzioni scolastiche. Segue un'analisi degli obblighi normativi, in particolare quelli previsti dall'AI Act e dalle disposizioni pertinenti del GDPR, che delineano ciò che è consentito e richiesto. Infine, si esaminano gli aspetti tecnici necessari per tradurre tali requisiti legali ed educativi in soluzioni operative. Questa struttura riflette l'approccio che molte istituzioni educative dovrebbero adottare nell'implementazione di strumenti basati sull'IA: partire dalla definizione degli obiettivi, proseguire con la verifica dei vincoli normativi, e infine procedere con l'implementazione o l'acquisizione di soluzioni tecnologiche adeguate.

Gli esempi presentati nel capitolo hanno valore illustrativo e sono di natura ipotetica, data la limitata disponibilità di casi pubblicamente documentati e validati in questo ambito. Essi sono stati scelti con l'obiettivo di esemplificare rischi ricorrenti e orientare lo sviluppo di buone pratiche.

[5] Regolamento sui servizi digitali; Regolamento sui mercati digitali , Legge sui dati, Legge sulla governance dei dati, Legge sulla sicurezza informatica e il Regolamento sulla resilienza informatica. La tabella 6 include una sintesi di queste leggi.

La prossima sezione introduce le principali nozioni educative, giuridiche e tecniche alla base della spiegabilità nell'IA. Successivamente, verrà mostrato come tali nozioni vengono applicate in tre casi ipotetici sull'uso di strumenti di valutazione automatizzata, tutoraggio intelligente e rilevamento dei contenuti generati dall'IA. Verranno analizzati gli aspetti educativi, giuridici e tecnici, presentate le sfide associate all'applicazione di tali strumenti e fornite raccomandazioni per i diversi stakeholder. La sezione si conclude con il riepilogo dei punti chiave e con alcune considerazioni su eventuali problemi di attuazione. La tassonomia delle dimensioni chiave della spiegabilità nei sistemi di IA per l'istruzione, riportata nella tabella 4, fornisce un ulteriore contesto di riferimento per il quadro normativo e sarà menzionata in tutte le sezioni.

2.2 Nozioni introduttive

Aspetti educativi

Quando si implementa la XAI in contesti educativi, è essenziale dare priorità e promuovere la trasparenza affinché i sistemi di IA possano spiegare in modo chiaro le decisioni prese, consentendo agli educatori e agli studenti di comprendere i risultati e ritenerli affidabili (Maity & Deroy, 2024). La selezione di modelli di intelligenza artificiale che bilancino adeguatamente le prestazioni con l'interpretabilità contribuisce in modo sostanziale a rafforzare la fiducia degli utenti nei sistemi automatizzati. Tuttavia, la trasparenza richiede che le spiegazioni siano specifiche rispetto al compito e orientate all'azione, affinché risultino effettivamente utili. Ad esempio, nel caso di un sistema automatizzato di valutazione, gli insegnanti hanno bisogno di spiegazioni locali, ossia riferite a singole decisioni, che permettano di comprendere il funzionamento del sistema e i criteri con cui vengono assegnati i voti. (Messer et al, 2024). Gli educatori devono fornire agli studenti spiegazioni non contraddittorie, descrivendo in dettaglio i fattori o le caratteristiche chiave considerati dall'IA, come la rubrica utilizzata, i suoi descrittori e le ponderazioni. Ciò garantisce che educatori e studenti possano fidarsi del ragionamento e dei risultati dell'IA.

Occorre inoltre definire chiaramente le responsabilità degli attori coinvolti, con processi che specifichino se siano gli sviluppatori, gli educatori o altri operatori a essere responsabili dei risultati dell'IA, e devono essere istituiti protocolli solidi di gestione degli errori per affrontare e correggere eventuali sbagli, garantendo che la supervisione umana rimanga presente e sia parte integrante del sistema. Gli sviluppatori dovrebbero essere responsabili della progettazione di algoritmi trasparenti, della riduzione al minimo dei bias e della fornitura di una documentazione dettagliata sul funzionamento del sistema. Gli educatori, in qualità di utenti finali, sono responsabili dell'interpretazione dei risultati dell'IA, della loro convalida rispetto al giudizio umano, dovendo altresì garantire che le raccomandazioni dell'IA siano in linea con gli obiettivi educativi. Gli operatori del sistema devono supervisionare il monitoraggio continuo delle prestazioni del sistema, risolvere gli errori e garantire il rispetto degli standard etici e di riservatezza.

Affrontare il problema dei bias e promuovere l'equità significa inoltre verificare come i sistemi di IA prendono le decisioni, ad esempio testando se gli studenti provenienti da contesti diversi ricevono feedback simili. Gli educatori possono supportare questo processo identificando pattern che il sistema potrebbe non rilevare.

Quando vengono rilevati bias, gli sviluppatori possono adeguare i dati o i metodi di valutazione di conseguenza. Per identificare efficacemente le distorsioni e promuovere l'equità, i decisori in campo educativo e gli operatori del sistema necessitano di spiegazioni su misura che soddisfino le esigenze specifiche dei soggetti interessati. Tali spiegazioni (Kim et al, 2024) promuovono la fiducia, l'equità e la trasparenza nei sistemi di IA. Ad esempio, le spiegazioni basate su approcci controfattuali mostrano perché un discente ha ottenuto un risultato diverso da un altro (Miller, 2018). Le spiegazioni basate sulle caratteristiche (Ribeiro et al, 2016) aiutano a identificare potenziali fonti di distorsioni, mentre le spiegazioni procedurali possono mostrare come è stata presa una decisione, il che è utile per le verifiche o la correzione di errori. In sintesi, spiegazioni chiare e personalizzate aiutano a comprendere e a valutare le decisioni dell'IA, ad affrontare le disparità e a mantenere la fiducia nei contesti educativi (Binns et al, 2018). Per poter correggere i bias e promuovere l'equità è necessario realizzare un'analisi dei dati di addestramento e dei risultati del sistema, al fine di identificare *pattern* da affrontare strategicamente, per esempio, con il ribilanciamento dei set di dati, l'adeguamento dei pesi algoritmici o l'integrazione di vincoli di equità. Ciò deve essere effettuato dagli sviluppatori durante la fase di convalida del sistema di IA. Altrettanto importante è una comunicazione trasparente su come i dati vengono raccolti, utilizzati e consultati. Gli educatori, gli studenti e i genitori dovrebbero comprendere quali tipi di dati utilizza il sistema di IA (ad esempio, i punteggi dei test o le metriche di partecipazione), in che modo questi dati influenzano le decisioni e chi ha accesso a tali dati. Il rispetto delle normative pertinenti, come il GDPR, garantisce la tutela della riservatezza e la protezione dei dati, mentre protocolli di consenso chiari e pratiche di anonimizzazione impediscono l'uso improprio di informazioni sensibili.

Lo sviluppo di politiche complete sull'uso della tecnologia negli istituti di istruzione comprende l'adozione di solide misure di protezione dei dati e della riservatezza, quali politiche dettagliate sul trattamento dei dati e rigorose misure di sicurezza informatica per salvaguardare le informazioni sensibili. L'uso etico della tecnologia dovrebbe essere guidato da un codice di condotta esplicito per prevenire abusi, quali il cyberbullismo o l'accesso non autorizzato ai dati. Servono infine politiche di regolamentazione che allineino l'uso etico dell'IA con i valori e gli obiettivi educativi dell'istituto (Paschal, 2023). Promuovere l'accessibilità e l'inclusività significa garantire a tutti gli studenti pari accesso alle tecnologie necessarie, comprese le tecnologie assistive, assicurando un ambiente di apprendimento inclusivo. Diversi esempi di utilizzo dei sistemi di IA nell'istruzione sono riportati nel rapporto sull'IA redatto dalla prima squadra EDEH sull'Intelligenza artificiale nell'istruzione https://data.europa.eu/doi/10.2797/828281 [versione in lingua italiana disponibile su sito Indire].

Lo sviluppo professionale degli educatori dovrebbe comprendere una formazione continua sull'integrazione efficace della tecnologia e sulla comprensione degli strumenti di IA, nonché garantire che tutte le parti interessate siano consapevoli e comprendano che politiche tecnologiche solide possono promuovere una cultura dell'uso etico ed efficace della tecnologia.

Potrebbero esserci diverse aree di interesse per integrare efficacemente l'IA nell'istruzione, che riguardano dimensioni diverse ma interconnesse di applicazione della tecnologia:

Uso trasformativo dell'IA (quadro SAMR): tale modello sottolinea il progresso dell'uso della tecnologia nella trasformazione delle pratiche educative. Analizzando come gli strumenti di IA possono migliorare, modificare o ridefinire le attività di apprendimento, gli educatori sono incoraggiati ad andare oltre la semplice sostituzione dei metodi tradizionali ed esplorare come l'IA possa migliorare radicalmente o creare nuove esperienze di apprendimento. Questa prospettiva è fondamentale per sfruttare le capacità uniche dell'IA, come il feedback personalizzato, i sistemi di apprendimento adattivo e la capacità di tali sistemi di modellare concetti complessi in modi precedentemente ritenuti impossibili.

Competenze interdisciplinari (quadro TPACK): un secondo modello evidenzia l'interazione tra le conoscenze tecnologiche, pedagogiche e contenutistiche di un educatore. Comprendere come questi ambiti si intersecano è fondamentale per progettare esperienze di apprendimento basate sull'IA che siano significative e pedagogicamente valide. Questo approccio garantisce che l'integrazione dell'IA non sia solo tecnicamente efficiente, ma anche in linea con i contenuti insegnati e le strategie utilizzate per trasmetterli. Esso dà priorità al ruolo dell'educatore nella creazione di lezioni che sfruttino efficacemente il potenziale dell'IA per migliorare la comprensione e il coinvolgimento degli studenti.

Sistemi e influenze esterne (quadro SETI): questo approccio adotta una prospettiva sistemica più ampia, riconoscendo che l'integrazione tecnologica è influenzata da una serie di fattori esterni che vanno oltre l'ambito scolastico. Tra questi figurano la disponibilità di infrastrutture, il sostegno istituzionale, le politiche e il contesto socioculturale. Tenendo conto di questi elementi, tale prospettiva garantisce che gli educatori non lavorino in modo isolato e che siano presenti i sistemi di supporto necessari, quali formazione, orientamento alla leadership e accesso equo agli strumenti di IA. Il quadro sottolinea inoltre l'importanza di affrontare gli atteggiamenti sociali e le norme culturali relative alla tecnologia, che possono influire in modo significativo sulla sua accettazione ed efficacia nei contesti educativi.

Insieme, le suddette aree di interesse possono evidenziare aspetti importanti da tenere in considerazione per un uso efficace dell'IA nell'istruzione e incoraggiare gli educatori a riflettere criticamente su come l'IA possa trasformare l'apprendimento, oltre a garantire un'integrazione della stessa che sia pedagogicamente fondata e sostenibile.

Aspetti legali

I diritti fondamentali sono profondamente radicati nel tessuto costituzionale dell'Unione Europea, fungono da fondamento basato sui valori per l'integrazione europea e forniscono un quadro normativo per l'agenda legislativa dell'UE[6]. L'approccio dell'UE alla regolamentazione del digitale si basa sui diritti fondamentali delle persone, ma al contempo vuole promuovere l'innovazione attraverso un'IA centrata sull'uomo e affidabile (Comunicazione della Commissione europea, 2018). In questo contesto, il regolamento sull'IA stabilisce norme per lo sviluppo, la commercializzazione e l'uso dell'intelligenza artificiale all'interno dell'Unione, integrando il GDPR e altre leggi sul digitale approvate dall'UE. Mentre il regolamento si applica ai sistemi di IA e ai modelli di IA per uso generale (GPAIM), il GDPR disciplina il trattamento dei dati personali.

[6] Articolo 2 del Trattato sull'Unione europea. Si veda inoltre la Convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto (2024) (si noti che il Consiglio d'Europa è indipendente dall'UE)

Quando un sistema di IA o GPAIM tratta dati personali, si applicano sia il GDPR che il regolamento sull'IA. Entrambe le normative si applicano a prescindere dal settore interessato.

Uno dei pilastri fondamentali del regolamento sull'IA è il suo quadro normativo basato sul rischio, che classifica i sistemi di IA in quattro **categorie** distinte: sistemi *vietati, ad alto rischio, a rischio minimo e a basso rischio*, in base ai potenziali rischi per gli individui e la società.

Portata del regolamento sull'IA (AI Act)

Il regolamento sull'IA disciplina due tipi di tecnologia, i sistemi di IA e i GPAIM:

Sistemi di IA[7] Qualsiasi sistema basato su macchine (software) progettato per operare con diversi livelli di autonomia al fine di inferire come generare output (quali previsioni, contenuti, raccomandazioni o decisioni) che possono influenzare ambienti fisici o virtuali, e che può continuare ad adattarsi anche dopo l'immissione sul mercato. *(Articolo 3(1), Considerando 12)*	**Modelli di IA per finalità generali (GPAIMs)** Un modello di IA che presenta una significativa generalità, è in grado di svolgere in modo competente un'ampia gamma di compiti distinti e può essere integrato in una varietà di sistemi o applicazioni a valle. *(Articolo 3(63), Considerandi 97, 98, 99)*

Ad esempio, uno strumento di intelligenza artificiale open source progettato per aiutare gli insegnanti nella valutazione dei compiti scritti potrebbe a prima vista sembrare esentato dagli obblighi previsti dal regolamento, ma poiché interagisce direttamente con i compiti consegnati dagli studenti e influenza le valutazioni individuali, dovrà rispettare i requisiti di trasparenza, equità e gestione dei rischi. Allo stesso modo, un GPAIM rilasciato da un'università per la ricerca sui metodi di apprendimento adattivo può beneficiare di alcune esenzioni ai sensi delle disposizioni sull'open source, ma se è integrato in un sistema commerciale di gestione dell'apprendimento per l'istruzione personalizzata, dovrà analogamente rispettare gli obblighi sanciti dal regolamento, compresi quelli volti a garantire l'accuratezza e la non discriminazione. Infine, un GPAIM che informa strategie educative su larga scala (ad esempio, decisioni di finanziamento) a causa del suo impatto sistemico non potrà beneficiare di esenzioni e dovrà quindi osservare le disposizioni relative ai sistemi di IA ad alto rischio (articolo 55, considerando 114, 115). Sono invece esentati dall'applicazione del regolamento i sistemi e i modelli di IA «sviluppati e messi in servizio specificamente al solo scopo di ricerca e sviluppo scientifico» (articolo 2, paragrafo 6, considerando 25); per esempio un'università che sviluppi uno strumento di tutoraggio basato sull'IA a fini di ricerca non sarà assoggettata al regolamento durante la fase di ricerca e sviluppo.

[7] Cfr. Commissione europea, Annex to the Communication to the Commission: Approval of the Content of the Draft Communication from the Commission – Commission Guidelines on the Definition of an Artificial Intelligence System Established by Regulation (EU) 2024/1689 (AI Act), C (2025) 924 final (6 febbraio 2025).

Termini chiave per comprendere la portata e l'applicabilità dell'AI Act

Spiegare cosa si intende per *"immissione sul mercato", "messa a disposizione sul mercato" e "messa in servizio"* è essenziale per comprendere la portata e l'applicabilità del regolamento, poiché tali termini definiscono le fasi critiche del ciclo di vita di un sistema di IA o di un GPAIM.

Immissione sul mercato (articolo 3, paragrafo 9)	La prima volta in cui un sistema di IA o un GPAIM è messo a disposizione per nell'UE. Ciò fa scattare i requisiti di conformità iniziali per i fabbricanti e gli sviluppatori.
Messa a disposizione sul mercato (articolo 3, paragrafo 10)	Fornitura di un sistema di IA o di un GPAIM per la distribuzione o l'uso nell'UE nel corso di un'attività commerciale, a titolo oneroso o gratuito. Ciò amplia il campo di applicazione per coprire l'intera catena di fornitura.
Messa in servizio (articolo 3, paragrafo 11)	la fornitura di un sistema di IA direttamente al deployer per il primo uso o per uso proprio nell'Unione per la finalità prevista. Tale fase segna l'inizio della diffusione operativa e mette in evidenza gli adempimenti da parte dei deployers e degli utenti finali.

Chi è assoggettato all'AI Act

Il regolamento identifica i principali soggetti (collettivamente denominati "operatori")[8]coinvolti nel ciclo di vita dell'IA:

Fornitori (articolo 3, paragrafo 3)	Sviluppatori di sistemi di IA o GPAIM, quali persone fisiche o giuridiche, autorità pubbliche o agenzie che immettono il sistema o modello sul mercato o in servizio con il proprio nome o marchio, a titolo oneroso o gratuito (ad esempio, società di tecnologia educativa, università e istituti di ricerca[9], agenzie governative o dipartimenti che sviluppano sistemi di IA per uso nell'istruzione pubblica), editori (editori educativi che creano piattaforme di contenuti basate sull'IA, libri di testo interattivi, generatori di quiz, ecc.) e fornitori di IA come servizio.
Deployers (articolo 3, paragrafo 4, considerando 13)	persone fisiche o giuridiche, autorità pubbliche, agenzie altri organismi che utilizzano un sistema di IA sotto la propria autorità, tranne nel caso in cui il sistema di IA sia utilizzato nel corso di un'attività personale non professionale. Tale definizione amplia l'ambito di applicazione per coprire l'intera catena di fornitura.
Importatori (articolo 3, paragrafo 6)	persone fisiche o giuridiche ubicate o stabilite nell'Unione che immettono sul mercato un sistema di IA recante il nome o il marchio di una persona fisica o giuridica stabilita in un paese terzo, anche in collaborazione con i fornitori.
Distributori (articolo 3, paragrafo 7)	persone fisiche o giuridiche, diverse dal fornitore o dall'importatore, che distribuiscono e gestiscono sistemi di IA in una catena di fornitura, o che mettono a disposizione un sistema di IA sul mercato dell'UE.

Chi è responsabile ai sensi dell'AI Act

Il regolamento si applica principalmente ai *fornitori* (ad esempio, gli sviluppatori di tecnologie didattiche che sviluppano o commissionano lo sviluppo di un sistema di IA o di un GPAIM), come, per esempio, le grandi aziende tecnologiche, i fornitori di servizi cloud e di infrastrutture, le comunità di IA open source e gli istituti di ricerca accademica. Tuttavia, sono previsti anche obblighi a carico dei *deployers*, ovvero delle persone fisiche o giuridiche che utilizzano o gestiscono un sistema di IA in un contesto professionale, come gli educatori, gli istituti di istruzione o gli operatori di sistemi. Pertanto, il *deployer* potrebbe essere una scuola che utilizza un sistema di IA per la valutazione automatizzata o un operatore di sistema presso un istituto di istruzione terziaria che gestisce strumenti per il monitoraggio del comportamento degli studenti attraverso l'uso di webcam e microfoni durante le verifiche online. Identificare la categoria cui appartiene ciascun soggetto interessato all'interno di un istituto è fondamentale per individuarne i diritti e gli obblighi, in particolare con riferimento ai sistemi di IA vietati e ad alto rischio.

Un leader educativo, come ad esempio un dirigente scolastico (deployer), che valuta l'acquisto di una soluzione in pronta consegna (ad esempio, una piattaforma di apprendimento personalizzato), avrebbe la necessità che il fornitore del sistema di IA fosse in grado di

[8] L'articolo 3, paragrafo 8, definisce "operatore" un fornitore, un fabbricante di prodotti, un utilizzatore, un rappresentante autorizzato, un importatore o un distributore

[9] Se il sistema è sviluppato esclusivamente a fini di ricerca, può essere esentato ai sensi dell'articolo 2, paragrafo 6. Tuttavia, se viene commercializzato o ampiamente diffuso, l'università diventa un fornitore ai sensi della legge sull'IA.

dimostrare la conformità del prodotto attraverso le apposite valutazioni e documentazioni. Il dirigente dovrebbe inoltre essere in grado di interpretare spiegazioni globali, complete e comparative delle funzionalità del sistema, al fine di verificare gli adempimenti del regolamento sull'IA da parte del fornitore e assicurarsi che la soluzione sia in linea con le politiche istituzionali.

Invece, un educatore, per esempio un docente (*deployer*) che utilizzi per il proprio corso di insegnamento un sistema di valutazione automatizzato, potrebbe aver bisogno di spiegazioni locali, selettive e condizionali per capire rapidamente e affrontare il motivo per cui uno studente specifico ha ricevuto un determinato voto. Tali spiegazioni puntuali e precise garantiscono la trasparenza e supportano un processo decisionale informato in vari ruoli educativi. È importante sottolineare che *la finalità prevista*[10] di un sistema di IA si riferisce all'uso specificato dal fornitore, compreso il contesto e le condizioni d'uso descritte nelle istruzioni del sistema, nel materiale promozionale e nella documentazione tecnica. Per conformarsi a queste specifiche, i *deployer* potrebbero fare affidamento su spiegazioni globali per comprendere il comportamento complessivo del sistema, le sue capacità e i suoi limiti in diversi scenari. Ad esempio, un istituto di istruzione superiore che utilizza un sistema di IA per individuare casi di disonestà accademica deve garantire il raggiungimento della finalità prevista di intercettare i comportamenti scorretti senza che gli studenti siano ingiustamente presi di mira per differenze linguistiche nello stile di scrittura o la partecipazione a pratiche collaborative legittime.

Le spiegazioni esaurienti inserite nella documentazione tecnica dovrebbero inoltre facilitare la comprensione più approfondita della progettazione e dei limiti del sistema, ad esempio descrivendo in dettaglio come un sistema di valutazione basato sull'intelligenza artificiale valuti i compiti nelle varie materie. Le spiegazioni condizionali chiariscono come il sistema opera in determinate condizioni – ad esempio, spiegando le regole che determinano l'attivazione di raccomandazioni per esercitazioni aggiuntive in una piattaforma di apprendimento personalizzato oppure la logica utilizzata da un sistema di monitoraggio delle presenze per segnalare le assenze. Integrando le suddette dimensioni di spiegabilità, i responsabili dell'implementazione possono garantire che il sistema funzioni in modo trasparente ed etico. Con spiegazioni specifiche al contesto, i *deployer* possono chiarire agli interessati (ad esempio, studenti o genitori) le decisioni prese dal sistema di IA. Questa trasparenza favorisce la fiducia e consente ai deployer di svolgere adeguatamente la loro funzione di sorveglianza, rispettando al contempo i diritti riconosciuti alle persone in materia di decisioni individuali automatizzate, compresa la profilazione ai sensi del regolamento sulla protezione dei dati personali (articolo 22, considerando 71 e 72 del GDPR).

[10] La finalità prevista è l'uso di un sistema di IA previsto dal fornitore, compresi il contesto e le condizioni d'uso specifici (articolo 3, paragrafo 12, del regolamento sull'IA)

Rischio inaccettabile: pratiche di IA vietate dall'AI Act[11] (Capo II, articolo 5)

Alcune applicazioni di intelligenza artificiale sono severamente vietate, in quanto rischiano di violare i diritti fondamentali e gli standard etici. Sono vietate pratiche quali *il social scoring (*articolo 5, paragrafo 1, lettera c), ossia l'uso di sistemi che classificano o valutano le persone, compresi ovviamente gli studenti e il personale scolastico, in base a tratti comportamentali o caratteristiche personali (ad esempio espressioni facciali o toni di voce), e altresì i sistemi di IA specificatamente immessi sul mercato e messi in servizio o utilizzati per dedurre le emozioni di una persona fisica nei luoghi di lavoro e negli istituti di istruzione (articolo 5, paragrafo 1, lettera f), considerando 44). Tali strumenti, progettati con la finalità specifica di rilevare - o utilizzati a tale scopo- le emozioni in classe durante le valutazioni o durante le interazioni tra educatori e studenti, sollevano serie preoccupazioni in materia di riservatezza[12] necessità del consenso e accuratezza delle interpretazioni. Tuttavia, lo stesso sistema di rilevamento delle emozioni (tecnologia di riconoscimento facciale) può essere applicato per scopi tra loro assai diversi: social scoring (vietato) o semplici metodi di verifica (a basso rischio). Questa dualità sottolinea l'importanza della regolamentazione contestuale. In questo scenario, la spiegabilità non riguarda tanto la determinazione della conformità intrinseca della progettazione di un sistema, quanto il contesto in cui viene messo in funzione. È quindi responsabilità del *deployer* [soggetto che utilizza il sistema] implementare il sistema, monitorarne l'uso previsto o la finalità garantendo l'osservanza della normativa vigente. Tuttavia, è bene sottolineare che i *deployer* di sistemi di riconoscimento delle emozioni devono informare le persone esposte a tali tecnologie, anche se l'utilizzo rientra tra gli altri scopi consentiti (Articolo 50, paragrafo 3, considerando 132, dell'AI Act).

La "spiegabilità" può aiutare a chiarire come viene utilizzato uno strumento e se viene utilizzato in modo etico. Il *deployer* dovrà quindi valutare diversi fattori-chiave di rischio, tra i quali:

1. **Inserimento dei dati:** quali informazioni relative agli utenti (finali) (ad esempio, studenti) vengono inserite nel sistema (ad esempio, voti, frequenza o modelli comportamentali)
2. **Previsione dei dati in uscita:** quali previsioni o decisioni vengono prese (ad esempio, raccomandare un tutoraggio o classificare gli studenti in gruppi di livello)
3. **Correlazione input-output:** in che modo i dati degli utenti (finali) influenzano le decisioni (per esempio, se la frequenza di uno studente venga ingiustamente correlata alla previsione delle sue abilità accademiche)

Come già detto, la spiegabilità non consiste semplicemente nel rendere interpretabili gli algoritmi, ma implica comprendere e comunicare il ruolo del sistema nel più ampio contesto in cui viene utilizzato.

[11] Vd. gli Orientamenti della Commissione sulle pratiche vietate di IA, pubblicati il 2 febbraio 2025, sei mesi dopo l'AI Act e scaricabili anche in lingua italiana dal link https://digital-strategy.ec.europa.eu/it/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

[12] I sistemi di rilevamento delle emozioni nei contesti educativi si scontrano con rilevanti vincoli legali ai sensi del GDPR, in particolare per quanto concerne la riservatezza, il consenso informato e l'accuratezza dei dati trattati. Le disposizioni chiave includono l'articolo 5 (che impone un trattamento dei dati lecito, trasparente e limitato alle finalità per le quali i dati sono stati raccolti; ciò significa che la raccolta di dati relativi alle espressioni facciali senza informare gli studenti o i genitori violerebbe la trasparenza); l'articolo 9 (che vieta l'usodi dati sensibili quali quelli sullo stato emotivo emozioni senza l'esplicito consenso dell'interessato). Inoltre, l'articolo 22 vieta i processi decisionali automatizzati che incidano significativamente sulla persona, come la profilazione degli studenti sulla base del rilevamento delle emozioni, senza un'adeguata sorveglianza da parte di esseri umani. Le istituzioni scolastiche sono tenute a effettuare valutazioni d'impatto sulla protezione dei dati (DPIA) ai sensi dell'articolo 35 per valutare i rischi e garantire la conformità del trattamento alle norme di legge. In fine, il consenso essere informato, specifico e revocabile, come indicato all'articolo 7 e ai sensi dell'articolo 25 la protezione dati deve essere garantita in via prioritaria fin dalla progettazione e per impostazione predefinita.

Sistemi di IA ad alto rischio (Capo III dell'AI Act)

Nel caso dei sistemi di IA ad alto rischio (articolo 6, paragrafo 2, allegato III, considerando 56, dell'AI Act), come i sistemi che determinano l'accesso agli istituti di istruzione e formazione professionale e quelli destinati a valutare i risultati dell'apprendimento o le ammissioni, gli istituti devono rispettare rigorose misure sulla trasparenza, sulla sorveglianza umana e sulla responsabilità. Tali disposizioni si applicheranno ai sistemi ad alto rischio di cui all'allegato III a partire dall'agosto 2026. La maggior parte degli obblighi di conformità ricade sui fornitori (ad esempio, gli sviluppatori di tecnologie per l'educazione). Ai sensi dell'articolo 6, paragrafo 3, della legge sull'IA, alcuni sistemi riferiti al settore ad alto rischio possono non essere considerati tali qualora i fornitori asseriscano in seguito ad autovalutazione, che tali sistemi non presentano rischi significativi per i diritti fondamentali o non influenzano materialmente i risultati del processo decisionale (considerando 53). Tuttavia, esistono anche obblighi di ampia portata per i *deployers* (ad esempio, dirigenti scolastici, educatori e altri operatori del settore educativo) che utilizzano questi sistemi in un contesto professionale. Tra questi, figura l'obbligo per un *deployer* di adottare misure tecniche e organizzative adeguate a garantire che il sistema di IA sia utilizzato in conformità alle istruzioni per l'uso che lo accompagnano, l'obbligo di monitorare il funzionamento del sistema e di attuare misure di sorveglianza umana competente nella misura in cui il *deployer* eserciti il controllo sul sistema. Per il *deployer* tale funzione di sorveglianza umana implica altresì (1) la garanzia che le misure pertinenti e adeguate di robustezza e cybersicurezza siano regolarmente monitorate per verificarne l'efficacia e vengano regolarmente adeguate o aggiornate; (2) la garanzia che i dati di input siano pertinenti e sufficientemente rappresentativi nella misura in cui il *deployer* esercita il controllo su tali dati; (3) la conservazione dei log generati automaticamente dal sistema di IA nella misura in cui siano sotto il suo controllo; (4) la consultazione dei rappresentanti dei lavoratori e l'informazione dei dipendenti interessati che saranno esposti a un sistema di IA ad alto rischio prima della messa in servizio o dell'uso di quest'ultimo sul luogo di lavoro; (5) informare le persone soggette all'uso di un sistema di IA ad alto rischio e del loro diritto a una spiegazione se il sistema viene utilizzato per prendere decisioni o assistere nel prendere decisioni su persone fisiche; (6) effettuare una valutazione dell'impatto del sistema nel contesto specifico del suo utilizzo.

Sorveglianza umana (articolo 14 dell'AI Act)

La funzione di "sorveglianza umana" o "human in the loop" (HITL), sopra descritta, comprende due aspetti dell'IA: lo sviluppo del sistema e la fase operativa. Il termine "loop" si riferisce infatti alle fasi del ciclo di vita del sistema di IA in cui la sorveglianza umana può rendersi necessaria per evitare rischi. In alcuni loop non richiedono l'HITL. Ad esempio, richiedono la sorveglianza umana i sistemi di IA che influenzano il progresso accademico, l'ammissione o la valutazione degli studenti (sistemi di IA ad alto rischio), mentre la sorveglianza non è necessariamente richiesta nel caso dei di IA che svolgono compiti di routine come l'automazione dei flussi di lavoro amministrativi. È quindi necessario delineare adeguatamente l'ambito di applicazione di questi loop per evitare una sorveglianza eccessiva. A tal fine, i *deployer* devono identificare il contesto e l'impatto che tali sistemi di IA hanno sul processo decisionale.

Ad esempio, in una scuola che utilizza sistemi di IA per formulare raccomandazioni di apprendimento, i cicli possono includere l'addestramento del modello, la sua integrazione nella piattaforma di apprendimento e le interazioni in tempo reale con gli studenti. All'interno di tali fasi, le interazioni in tempo reale possono richiedere la sorveglianza HITL, mentre per le interazioni più semplici saranno sufficienti revisioni periodiche.

Inoltre, è fondamentale scegliere l'esperto giusto per l'HITL. Le competenze necessarie dipendono dallo scopo della sorveglianza. Ad esempio, se l'obiettivo è l'accuratezza nella valutazione degli elaborati, la persona cui è affidata la sorveglianza dovrebbe essere esperta nella materia considerata e comprendere i metodi di valutazione dello strumento di IA. Se invece l'obiettivo è garantire l'equità nelle decisioni di ammissione, l'essere umano che sorveglia dovrà comprendere i principi di giustizia ed equità, nonché le metriche utilizzate dal sistema di IA. La spiegabilità aiuta a definire gli obiettivi della sorveglianza HITL. Chiarendo cosa fa il sistema e come lo fa, le istituzioni possono adeguare meglio i ruoli di sorveglianza ai rischi o agli obiettivi specifici associati all'uso dell'IA. Ad esempio, se il sistema di IA è progettato per prevedere il successo accademico, la spiegabilità aiuta a determinare se l'essere umano che lo sorveglia ha bisogno di competenze per l'analisi dei dati accademici, in materia di orientamenti etici o di politiche istituzionali. È importante notare che una soluzione HITLè valida solo quanto il ciclo in cui è inserita. Per svolgere adeguatamente la funzione di sorveglianza umana, è essenziale definire chiaramente il ciclo, comprendere perché la sorveglianza è necessaria e mettere in atto un processo per eliminare i rischi sistemici. Senza tali accorgimenti l'approccio HITL non funzionerà.

Il ruolo della spiegabilità e dell'HITL diventa ancora più centrale con riferimento ai sistemi biometrici utilizzati in contesti educativi[13]. Sistemi biometrici, come il riconoscimento facciale, vengono spesso utilizzati per monitorare la presenza, migliorare la sicurezza nei campus, nelle aule o sale d'esame, o per identificare le persone durante grandi raduni studenteschi o su piattaforme di apprendimento online. Queste applicazioni possono essere utili, ma devono essere implementate con cautela. I sistemi biometrici utilizzati esclusivamente a fini di verifica, come i meccanismi di login per confermare l'identità dello studente quando accede a risorse o piattaforme sono esclusi dalla classificazione come sistemi ad alto rischio. Tuttavia, chi li impiega deve comunque informare gli interessati del fatto che su di loro vengono applicati sistemi di categorizzazione biometrica (articolo 50, paragrafo 3, dell'AI Act). Inoltre, applicazioni di più ampia portata, come il tracciamento o la vigilanza, sono considerate ad alto rischio a causa del potenziale abuso che si verificherebbe in caso di tracciamento non autorizzato di studenti o personale, violazioni della privacy o condivisione di dati biometrici senza un consenso adeguato.

I responsabili del trattamento che utilizzano sistemi di categorizzazione ad alto rischio devono informare le persone fisiche esposte a tali sistemi (articolo 50, paragrafo 3, considerando 132, dell'AI Act) e trattare i dati da essi generati in conformità al GDPR. La trasparenza è essenziale e le istituzioni hanno l'obbligo di comunicare la finalità della raccolta e della conservazione di tali dati biometrici. Inoltre, agli studenti e al personale devono essere forniti meccanismi di rinuncia agli usi non essenziali, garantendo così la tutela dei loro diritti e promuovendo un clima di fiducia all'interno dell'ambiente educativo.

[13] Il «sistema di categorizzazione biometrica ad alto rischio» di cui all'articolo 6, paragrafo 2, e definito nell'allegato III, si riferisce ai sistemi utilizzati per finalità quali l'identificazione di caratteristiche sensibili o protette, in quanto tali sistemi possono causare danni significativi o influenzare i risultati del processo decisionale.

La spiegabilità garantisce la trasparenza nel funzionamento di tali sistemi, consentendo ai *deployer* e al personale di sorveglianza di comprendere come i dati biometrici vengono raccolti, trattati e utilizzati e se ciò è in linea con la finalità prevista del sistema. Ad esempio, la spiegabilità consente alla persona che sorveglia un sistema biometrico di rilevazione delle presenze di verificare che i dati raccolti siano utilizzati esclusivamente a fini di rilevazione delle presenze e non per il tracciamento o la profilazione non autorizzati.

Per i sistemi ad alto rischio come i biometrici è previsto l'obbligo di valutazione d'impatto sui diritti fondamentali (FRIA) (articolo 27, considerando 93 e 96, dell'AI Act). I *deployer* di sistemi ad alto rischio, compresi gli enti pubblici e i soggetti privati che forniscono servizi pubblici, devono completare una FRIA prima di implementare tali sistemi. Tale impegno comporta l'identificazione delle persone interessate, la valutazione dei rischi d'impatto sui diritti fondamentali e l'attuazione di strategie di sorveglianza e attenuazione dei rischi. Il processo FRIA migliora la comprensione del sistema di IA e dei suoi dati, promuovendo la trasparenza per le parti interessate e fornendo un quadro di riferimento per integrare la spiegabilità e l'HITL nel funzionamento dei sistemi di IA ad alto rischio.

Sistemi di IA a rischio minimo di cui all'articolo 50 dell'AI Act (IA rivolta a singoli utenti)

Gli obblighi di trasparenza per determinati sistemi di IA sono stabiliti all'articolo 50 della legge dell'AI Act. I fornitori devono informare le persone fisiche che li utilizzano che stanno interagendo con sistemi di IA, come i chatbot. Gli output di IA generativa, sia sotto forma di contenuti audio, immagine, video o testuali (ad esempio deepfake), devono essere marcati in un formato leggibile meccanicamente e rilevabili come generati o manipolati artificialmente. Tale requisito è fondamentale nei contesti educativi, dove l'IA generativa potrebbe essere utilizzata per creare materiali didattici, feedback o comunicazioni. Un'etichettatura chiara contribuisce a mantenere la fiducia e a prevenire gli abusi.

Diritto alle spiegazioni

Le persone interessate hanno anche il diritto di ricevere spiegazioni chiare e significative dal responsabile del trattamento in ordine alle modalità di coinvolgimento del sistema di IA nel processo decisionale (articolo 86, dell'AI ACT). Tali obblighi integrano i principi di protezione dei dati[14] in materia di trasparenza, come la necessità generale di un diritto alla comunicazione trasparente (articoli da 12 a 14 del GDPR), che impone agli istituti di istruzione (in qualità di titolari del trattamento) di fornire informazioni sulle attività di trattamento in forma «concisa, trasparente, intelligibile e facilmente accessibile, con un linguaggio semplice e chiaro, in particolare nel caso di informazioni destinate specificamente ai minori»[15] Gli studenti e le altre persone interessate hanno il diritto di essere informati quando, ad esempio, nei contesti educativi vengono utilizzati processi decisionali automatizzati, quali la valutazione algoritmica o i sistemi di apprendimento personalizzati. Inoltre, gli interessati hanno il diritto di accedere alle informazioni relative alle decisioni automatizzate che li riguardano, compresi i dettagli sulla logica decisionale e le implicazioni della decisione per la loro esperienza educativa (articolo 15, considerando 63 e 71, GDPR). Gli interessati hanno infine il diritto di opporsi al trattamento dei propri dati personali, compresa la profilazione (articolo 21, GDPR).

[14] Articolo 5, GDPR: (1) liceità, correttezza e trasparenza; (2) limitazione delle finalità; (3) minimizzazione dei dati; (4) accuratezza; (5) limitazione della conservazione; (6) integrità e riservatezza.
[15] Una dichiarazione sulla privacy dovrebbe essere collegata in fondo a ogni pagina del sito web. Una dichiarazione sulla privacy a più livelli dovrebbe essere breve, concisa e completa. Cfr. Gruppo di lavoro articolo 29, parere 10/2004.

Tale diritto assume particolar e rilevanza nel settore dell'istruzione, quando la profilazione è utilizzata per finalità di monitoraggio dei risultati scolastici o per la previsione del comportamento degli studenti. Ai fini del marketing diretto, come la promozione di servizi aggiuntivi agli studenti, questo diritto è assoluto. Infine, l'articolo 35 impone agli istituti di istruzione di effettuare valutazioni d'impatto sulla protezione dei dati (DPIA) prima di implementare sistemi di IA che trattano dati personali in modalità che possano comportare un rischio elevato per i diritti e le libertà delle persone. Ad esempio, le DPIA sono necessarie per valutare i sistemi automatizzati utilizzati nelle procedure di ammissione, nella valutazione e nel supporto agli studenti, poiché questi processi possono avere un impatto significativo sul percorso formativo di questi ultimi.

La mappa di conformità XAI-Ed, riportata di seguito (tabella 5), delinea gli obblighi specifici in capo ai deployer e ai fornitori, allineando i sistemi di IA per l'istruzione ai principi XAI. Questa mappatura vuole fungere da strumento di riferimento di alto livello.

Sistemi di IA ad alto rischio: Capo III, AI Act		
Obblighi fondamentali	**Fornitore**	**Deployer**
Sistemi di gestione dei rischi Articolo 9	Implementare un sistema di gestione dei rischi relativi a distorsioni, equità e trasparenza negli strumenti di IA.	-
Dati e governance dei dati Articolo 10	Stabilisce i requisiti per l'addestramento, la convalida e la prova dei set di dati. Devono essere pertinenti, sufficientemente rappresentativi, privi di errori e completi in relazione alla finalità prevista.	-
Documentazione tecnica Articolo 11	Stabilisce le norme per la redazione di documenti tecnici sui sistemi di IA ad alto rischio prima della loro immissione sul mercato.	-
Conservazione dei dati Articolo 12	Stabilisce le norme per la registrazione automatica degli eventi (log), durante il ciclo di vita di un sistema di IA.	-
Trasparenza e fornitura di informazioni ai deployer Articolo 13	I sistemi di IA devono essere progettati in modo da essere trasparenti, affinché i deployer [utilizzatori] possano comprenderli e utilizzarli correttamente. Le istruzioni devono essere chiare e includere informazioni sul fornitore, sulle capacità, sui limiti del sistema e sui rischi. Devono spiegare come interpretare gli output del sistema, eventuali modifiche predeterminate al sistema e come mantenerlo. Le istruzioni devono descrivere come raccogliere, conservare e interpretare i registri dei dati.	Condividere spiegazioni comprensibili con gli utenti finali (ad esempio studenti, genitori e personale).
Sorveglianza umana Articolo 14	Progettare sistemi che permettano una efficace supervisione umana. Tali misure dovrebbero essere commisurate al rischio, adeguate al contesto ed integrate nel sistema dal fornitore. I sistemi di IA devono includere meccanismi per guidare e informare la persona cui è affidata la sorveglianza affinché possa prendere decisioni informate su quando e come intervenire.	Garantire un controllo efficace dei sistemi di IA utilizzati nelle operazioni volte a prevenire o ridurre al minimo i rischi in base alle finalità previste o agli usi impropri ragionevolmente prevedibili.

Accuratezza, robustezza e cibersicurezza Articolo 15	Progettare sistemi di IA robusti per l'istruzione con meccanismi per gestire le imprecisioni e i risultati distorti. Proteggere i sistemi di IA da attacchi di terzi non autorizzati.	-
Obblighi dei fornitori Articolo 16	Rispettare i requisiti relativi alla valutazione di conformità e conservare la documentazione. Conformità alla marcatura CE.	-
Obblighi del deployer Articolo 26	-	Obbligo di adottare misure tecniche e organizzative adeguate e di affidare la sorveglianza umana, ad esempio attuando misure volte a garantire un uso sicuro ed equo dell'IA in contesti educativi.
Valutazione di impatto sui diritti fondamentali (FRIA) Articolo 27	-	I deployer che sono organismi di diritto pubblico o soggetti privati che forniscono servizi pubblici devono valutare l'impatto sui diritti fondamentali che l'uso del sistema adalto rischio può produrre.
Monitoraggio post-commercializzazione Articolo 72	Monitorare le prestazioni degli strumenti di IA dopo la messa in servizio.	Segnalare i problemi e verificare l'efficacia del sistema.
Segnalazione di incidenti gravi Articolo 73	Segnalare qualsiasi incidente grave all'autorità di vigilanza del mercato entro i termini specificati; presentare una relazione; indagare tempestivamente sull'incidente, individuarne la causa principale e collaborare con le autorità competenti per garantire la risoluzione e prevenire il ripetersi dell'incidente.	Istituire meccanismi per monitorare e individuare gli incidenti gravi; segnalare i casi sospetti al fornitore e, se necessario, alle autorità competenti.
Diritto alla spiegazione Articolo 86	-	Qualsiasi persona interessata da una decisione del deployer ha il diritto di ottenere una spiegazione "chiara e significativa" dal deployer (considerando 171). Ad esempio, gli studenti e i genitori hanno il diritto di ottenere informazioni sulle decisioni prese dall'IA.
Sistemi di IA a basso rischio		
Obblighi fondamentali	**Fornitore**	**Deployer**
Obblighi di trasparenza dei fornitori e dei deployers di determinati sistemi di IA Articolo 50	I fornitori devono informare gli utenti che stanno interagendo con un sistema di IA, ad esempio una chatbot o un sistema di riconoscimento delle emozioni, o che stanno visualizzando i risultati di sistemi di IA, ad esempio i deepfake. I sistemi di IA che creano contenuti, compresi i sistemi di IA generici, garantiscono che i propri output siano marcato in un formato leggibile da una macchina.	I deployer che utilizzano un sistema di riconoscimento delle emozioni o di categorizzazione biometrica devono informare le persone sul funzionamento del sistema e sul trattamento dei loro dati in conformità con gli obblighi in materia di protezione dati (GDPR).
Obblighi di trasparenza per i fornitori e gli utenti di determinati sistemi di IA Articolo 50 **Codici di condotta per l'applicazione volontaria di requisiti specifici Articolo 95**	L'Ufficio dell'UE per l'IA e gli Stati membri incoraggiano l'elaborazione di codici di condotta per i sistemi di IA. Tali codici promuoveranno l'adesione volontaria a determinati standard, tenendo conto delle soluzioni tecniche e delle migliori pratiche del settore.	I deployer possono decidere di seguire codici di condotta per l'applicazione volontaria di requisiti specifici.

Tabella 5: Mappa di conformità XAI-Ed.

Oltre all'AI Act e al GDPR, esistono svariate disposizioni UE sul digitale che sono rilevanti per il settore dell'istruzione.

La tabella 6 riportata di seguito fornisce una panoramica della normativa vigente o di prossima applicazione.

Normativa	Area di interesse	Rilevanza per l'istruzione	Collegamento con la spiegabilità
Regolamento sui servizi digitali (DSA) (entrato in vigore il 17 febbraio 2024 e applicabile alle piattaforme online di dimensioni molto grandi (VLOP) e ai motori di ricerca online di dimensioni molto grandi (VLOSE) dal 25 agosto 2023)	Trasparenza degli algoritmi delle piattaforme, dei diritti degli utenti e della moderazione dei contenuti.	Trasparenza negli algoritmi utilizzati per le piattaforme di apprendimento online (ad esempio, selezione dei contenuti, moderazione e sistemi di raccomandazione). L'articolo 28 garantisce la protezione dei minori. I VLOP e i VLOSE (ad esempio YouTube e Google) devono proteggere i dati degli utenti e limitare i contenuti illegali/inappropriati. Vieta la pubblicità mirata ai minori e l'utilizzo di dati personali sensibili.	Obbliga le piattaforme a fornire spiegazioni chiare sul funzionamento degli algoritmi con riguardo alla selezione e alla moderazione dei contenuti. Aiuta educatori e studenti a comprendere i processi alla base della raccomandazione di contenuti e la moderazione delle classi online. I ricercatori hanno accesso ai dati delle piattaforme chiave per esaminarne il funzionamento. Obblighi di comunicazione trasparente per i servizi intermediari, i servizi di memorizzazione delle informazioni (hosting), le piattaforme online e VLOP.
Regolamento sui mercati digitali (DMA) (entrato in vigore il 1° novembre 2022)	Concorrenza leale e portabilità dei dati nei mercati digitali.	Si applica alle grandi aziende tecnologiche (gatekeeper come Alphabet, Amazon, Apple, Meta, Microsok, ecc.) Garantisce un accesso equo alle piattaforme e agli strumenti didattici. Impone la portabilità dei dati per gli istituti di istruzione che cambiano piattaforma (ad esempio, passando da un sistema di gestione dell'apprendimento a un altro).	Obbliga i gatekeeper a chiarire le modalità di trattamento e alla conservazione dei dati da parte delle piattaforme. Facilita l'interoperabilità garantendo la trasparenza del trattamento dei dati quando gli istituti cambiano piattaforma.
Regolamento sui dati (entrato in vigore l'11 gennaio 2024. Si applicherà a partire dal 12 settembre 2025)	Garantisce la condivisione sicura dei dati e l'interoperabilità, in particolare per i dati non personali.	Incoraggia l'innovazione nell'edtech consentendo alle istituzioni e ai ricercatori di accedere in modo sicuro e comprendere i dati educativi non personali. Stabilisce chi può utilizzare quali dati e a quali condizioni. Sottolinea l'importanza dell'alfabetizzazione dei dati.	Garantisce che i responsabili del trattamento forniscano spiegazioni chiare sul trattamento dei dati, consentendo analisi accurate ed eque a fini di ricerca o formazione.
Regolamento sulla governance dei dati (DGA) (entrato in vigore il 23 giugno 2022, applicabile dal settembre 2023)	Trasparenza nei meccanismi di condivisione dei dati e intermediari affidabili.	Promuove meccanismi trasparenti di condivisione dei dati per gli istituti di istruzione. Consente alle università di sviluppare programmi di studio basati sull'intelligenza artificiale utilizzando set di dati condivisi, comprendendo al contempo come i dati influenzano il processo decisionale.	Gli intermediari affidabili devono fornire informazioni chiare sul trattamento e la condivisione dei dati, promuovendo la trasparenza nelle applicazioni educative basate sull'IA.
Regolamento sulla cibersicurezza **(entrato in vigore il** 27 giugno 2019**)**	Certificazione di sicurezza per i fornitori e i sistemi ICT.	Migliora la sicurezza informatica degli ambienti e degli strumenti di apprendimento digitale. Consente alle istituzioni di valutare le misure di sicurezza degli strumenti certificati, garantendo la protezione dei dati degli studenti e del personale.	Sensibilizzazione alla cibersicurezza e promozione dell'alfabetizzazione informatica negli istituti di istruzione. I sistemi di certificazione richiedono ai fornitori di documentare e comunicare i protocolli di sicurezza, garantendo una chiara comprensione delle misure di protezione.
Regolamento sulla ciberresilienza (CRA) (entrato in vigore il 10 dicembre 2024)	Principi di sicurezza intrinseca per i dispositivi connessi e il software. *Completa la direttiva NIS2 2022*	Incoraggia l'innovazione nell'**edtech** consentendo alle istituzioni e ai ricercatori di accedere in modo sicuro e comprendere i dati educativi non personali. Regola chi può utilizzare quali dati e a quali condizioni. *Sottolinea l'importanza dell'alfabetizzazione dei dati.*	Garantisce che i responsabili del trattamento forniscano spiegazioni chiare sul trattamento dei dati, consentendo analisi accurate ed eque a fini di ricerca o formazione.

Tabella 6: Panoramica della normativa UE sul digitale rilevante per il settore dell'istruzione.

Aspetti tecnici

La dimensione tecnica della progettazione e dell'implementazione di soluzioni spiegabili per gli attori del settore educativo non si limita alle sfide proprie dello sviluppo tecnologico del software, ma implica anche la necessità di bilanciare le pratiche consolidate nella realizzazione di sistemi di IA complessi con un approccio centrato sull'utente, orientato a garantire la massima trasparenza e interpretabilità. I fornitori di sistemi di IA sono tenuti a includere una documentazione chiara, redatta in un linguaggio accessibile. Ciò implica che le istruzioni siano concise, complete, corrette e comprensibili, offrendo informazioni pertinenti, facilmente interpretabili e utili per insegnanti, studenti e altri stakeholder. Strumenti di visualizzazione, come dashboard e indicatori di progresso, possono facilitare la comprensione degli *insight* generati dai dati e delle prestazioni del sistema, rendendoli più facilmente interpretabili (si veda il prossimo capitolo). Un ulteriore aspetto cruciale del processo di sviluppo software, al fine di garantire sistemi di IA spiegabili ed efficaci in ambito educativo, consiste nella validazione delle soluzioni tecniche all'interno del contesto educativo di riferimento, coinvolgendo gli stakeholder pertinenti. Questo passaggio è essenziale per instaurare una comunicazione efficace e per adattare le tecniche di *Explainable AI* (XAI) alle specificità di ciascun caso d'uso.

Nel campo dell'istruzione, il ruolo della XAI va oltre le sfide tecniche per rispondere alle diverse esigenze degli stakeholder, tra i quali studenti, educatori, amministratori e soggetti giuridici. I sistemi di IA applicati all'istruzione sono spesso costituiti da strutture complesse in cui collaborano più modelli di IA, richiedendo un livello di spiegabilità che vada oltre i risultati dei singoli modelli. Una XAI efficace deve fornire spiegazioni trasparenti e comprensibili delle decisioni prese dall'IA, adattate alle esigenze di ciascun soggetto interessato, come verrà illustrato in dettaglio nel prossimo capitolo.

Da un punto di vista tecnico, lo sviluppo della XAI nell'istruzione richiede una solida comprensione delle esigenze degli stakeholder, che devono essere tradotte in requisiti tecnici espliciti a garanzia della responsabilità e della legalità.

Tale processo comporta una serie di sfide, tra cui l'integrazione di stakeholder eterogenei nella progettazione della XAI, la gestione dei differenti livelli di alfabetizzazione digitale e conoscenza dell'IA, nonché la considerazione degli aspetti pedagogici e giuridici. Un elemento cruciale è la definizione chiara degli utenti finali, poiché da essa dipende la selezione degli algoritmi e dei formati esplicativi più appropriati (ad esempio: spiegazioni testuali, visuali, basate su caratteristiche o su esempi). Le tecniche di XAI adottate devono essere coerenti con lo specifico caso d'uso e con le esigenze degli stakeholder coinvolti. A titolo esemplificativo, possono risultare più efficaci spiegazioni globali per i decisori politici, al fine di comprendere il comportamento complessivo del sistema, mentre spiegazioni locali sono più adatte per gli studenti, che necessitano di chiarimenti puntuali sui risultati individuali.

Gli stakeholder del settore tecnologico spesso non dispongono delle conoscenze necessarie in materia di istruzione per progettare contenuti e formati ottimali delle spiegazioni relative all'IA. Pertanto, sono necessari approcci di co-progettazione che utilizzino interfacce e canali di comunicazione tra esperti di materie diverse. A tal fine, la progettazione XAI può essere supportata da (1) definizioni chiare dei termini e del vocabolario utilizzati nelle discipline corrispondenti e (2) elenchi chiari dei requisiti, delle funzioni e delle caratteristiche definite per il caso d'uso. Attraverso la comunicazione, gli sviluppatori possono aiutare gli stakeholder del settore dell'istruzione a tradurre i requisiti pedagogici in funzioni e caratteristiche tecniche del sistema XAI.

Cosa spiegare e come spiegare

Le spiegazioni tecniche dei modelli di IA si concentrano sui meccanismi che hanno portato alla generazione di una struttura di previsione, sulle prestazioni e sui dati di addestramento, ma gli educatori necessitano di informazioni più approfondite sulle ipotesi di progettazione, sul ragionamento e sulle relazioni input-output dei modelli. Nelle attuali discussioni sulla XAI, l'attenzione è rivolta alla spiegazione dei modelli stessi di IA e non tanto all'ambiente in cui essi sono stati sviluppati, che comprende le ipotesi di progettazione, i principi di raccolta dei dati, l'interpretazione dei dati, l'etichettatura dei dati di addestramento e altri servizi connessi come l'hosting dei modelli. Ad esempio, quando un modello utilizza i movimenti oculari per tracciare lo sguardo, un educatore potrebbe aver bisogno di chiarimenti su come i dati relativi ai movimenti oculari siano correlati al rilevamento della presenza sullo schermo, che deriva da ipotesi formulate durante la preparazione dei dati e non solo dal funzionamento tecnico del modello. In altre parole, ciò amplia la spiegabilità del modello spostandosi sulla trasparenza del processo, che richiede alle soluzioni di IA spiegazioni comprensibili e orientate alla pedagogia. I sistemi educativi assistiti dall'IA devono dimostrare flussi di dati chiari e fornire un solido supporto per l'audit.

Gli sviluppatori possono scegliere tra una serie di tecniche XAI (Bennetot, 2024). Sebbene la scelta delle modalità di spiegazione dipenda in larga misura dallo specifico caso d'uso, esistono requisiti normativi vincolanti su ciò che i sistemi di intelligenza artificiale devono rendere esplicito, i quali includono, a titolo esemplificativo ma non esaustivo, quelli di documentazione e trasparenza ai sensi degli articoli 11 e 13 del capo III del regolamento sull'AI, riguardante i sistemi ad alto rischio. Tra gli altri requisiti legali, ci si attende che gli sviluppatori monitorino il livello di rischio del sistema che forniscono e siano obbligati a soddisfare i requisiti di spiegabilità, trasparenza e comprensibilità..

Per generare una spiegazione sul funzionamento interno di un sistema di IA, a partire dagli approcci introdotti nella sezione 1.1., gli sviluppatori utilizzano solitamente un approccio "post-hoc" per spiegare i modelli black-box, mentre i modelli open-box vengono spiegati utilizzando un approccio "ante-hoc":

Metodi di spiegazione "post-hoc: i metodi post-hoc mirano a fornire spiegazioni relative a modelli di tipo black-box una volta che questi sono stati sviluppati, senza modificarne la struttura interna. Tali approcci consentono di ottenere indicazioni sul funzionamento del modello e sulle ragioni alla base delle decisioni prodotte, contribuendo così a una maggiore comprensione e trasparenza del processo decisionale algoritmico.

Tecniche di rilevanza delle caratteristiche: valutano l'influenza delle singole caratteristiche sulle predizioni del modello.

- SHAP (Shapley additive explanations): utilizza la teoria dei giochi per assegnare punteggi di importanza alle caratteristiche, garantendo risultati coerenti e interpretabili in varie combinazioni di caratteristiche.
- LIME (spiegazioni locali interpretabili indipendenti dal modello): crea un modello locale interpretabile attorno a una previsione specifica perturbando i dati di input e analizzando i cambiamenti nei risultati.
- Analisi della sensibilità delle caratteristiche: misura l'impatto della modifica delle caratteristiche di input sui risultati del modello, identificando i fattori più influenti nel processo decisionale.

Spiegazioni controfattuali: forniscono scenari ipotetici, mostrando come la modifica di determinate caratteristiche potrebbe portare a risultati diversi. Ad esempio, in un sistema che prevede l'abbandono scolastico degli studenti, potrebbero indicare che una minore frequenza alle lezioni comporta una maggiore possibilità di abbandono.

Spiegazioni visive: tecniche come le mappe di salienza e Grad-CAM identificano le regioni dei dati di input (ad esempio, pixel di immagini o segmenti di testo) che influenzano fortemente le previsioni. I metodi di riduzione della dimensionalità, come PCA e t-SNE, semplificano i dati ad alta dimensionalità per la visualizzazione, aiutando a identificare i modelli chiave nei dati.

Spiegazioni tramite semplificazione: modelli semplificati, come gli alberi decisionali, approssimano il comportamento di modelli complessi per renderne comprensibile la logica.
Spiegazioni tramite esempi: consiste nel mostrare esempi reali o sintetici che illustrano le decisioni del modello, come la visualizzazione di un sottoinsieme di immagini classificate con un'etichetta specifica.

Metodi di spiegazione ante hoc: I metodi *ante-hoc* sono intrinsecamente interpretabili, in quanto progettati per garantire trasparenza già nella fase di costruzione del modello. Tali approcci permettono agli utenti di comprendere il processo decisionale senza la necessità di ricorrere a ulteriori algoritmi esplicativi. Tuttavia, anche quando il modello è tecnicamente trasparente, è fondamentale che la spiegazione venga tradotta in un formato comprensibile per gli stakeholder non tecnici. Alcuni esempi di metodi XAI ante hoc ben noti sono:

Alberi decisionali: questi modelli gerarchici suddividono i dati in rami in base ai valori delle caratteristiche, offrendo spiegazioni chiare e dettagliate dalla radice all'output.

Regressione lineare e logistica: questi modelli utilizzano coefficienti caratteristici per mostrare direttamente in che modo ciascuna variabile contribuisce alle previsioni. Ad esempio, in un sistema di rilevamento dell'abbandono scolastico, potrebbero mostrare che risolvere i compiti in ritardo nel semestre è associato a tassi di abbandono più elevati.

Modelli additivi generalizzati (GAM): i GAM consentono relazioni non lineari tra singole caratteristiche e risultati, mantenendo l'interpretabilità. È possibile visualizzare il contributo di ciascuna caratteristica, bilanciando complessità e trasparenza.

I metodi XAI non vengono implementati in modo isolato, ma all'interno di quadri concettuali più ampi che considerano l'intero sistema, comprendendo il modello, l'infrastruttura tecnologica e le interazioni con gli utenti. Tali framework (Khosravi et al, 2022; Mohseni, 2019; Liao et al, 2020) pongono particolare enfasi sull'integrazione del paradigma *human-in-the-loop* (HITL), assicurando che l'expertise umana sia coinvolta non solo nella fase di valutazione della spiegabilità, ma anche nella progettazione e nello sviluppo dei sistemi di IA. Gli approcci HITL danno priorità alla produzione di spiegazioni contestualmente rilevanti e coerenti con gli obiettivi educativi, migliorando così i processi decisionali grazie alla combinazione tra la precisione dell'IA e il giudizio umano. Questo approccio collaborativo garantisce che le predizioni generate dai sistemi di IA e le relative spiegazioni siano eticamente fondate, pedagogicamente appropriate e adeguate ai bisogni specifici degli utenti. Per rispondere in modo efficace alle esigenze eterogenee presenti nei contesti educativi, risultano fondamentali modelli di spiegabilità dinamici e centrati sugli stakeholder, strutturati su livelli multipli di dettaglio. Processi di co-progettazione continuativa e meccanismi iterativi di raccolta del feedback da parte degli utenti contribuiscono a perfezionare tali modelli, assicurando che essi rimangano chiari, utilizzabili e pertinenti nei diversi ambienti educativi.

2.3 Scenari di utilizzo

Strumenti di rilevamento dei contenuti AI

Emil è uno studente sedicenne dell'ultimo anno delle superiori. Stava studiando molto per mantenere buoni voti e prepararsi all'ingresso all'università. Lavorava part-time, studiava intensamente e partecipava ad attività extracurriculari. Nell'ambito del corso di storia, ha completato un progetto di ricerca che rappresentava una parte significativa della valutazione finale. Tuttavia, la settimana successiva alla consegna, è rimasto sorpreso nell'apprendere di aver fallito la valutazione digitale. Alla richiesta di spiegazioni rivolta all'insegnante, gli è stato riferito che uno strumento di rilevamento dell'intelligenza artificiale aveva segnalato il suo elaborato come probabilmente generato da un sistema automatizzato. Lo stesso strumento aveva inoltre evidenziato due precedenti compiti di Emil per le medesime ragioni.

Aspetti educativi

Questo caso evidenzia importanti questioni relative all'equità e all'uso etico dell'intelligenza artificiale in ambito educativo. Sebbene possano risultare utili, gli strumenti di rilevamento dell'IA non sono infallibili. I modelli probabilistici su cui si basano tali strumenti possono infatti identificare erroneamente stili di scrittura eterogenei, in particolare quelli adottati da studenti non di madrelingua (italiana) o con differenti abilità. Attualmente, gli strumenti preposti alla rilevazione di contenuti generati da intelligenze artificiali presentano un elevato tasso di errore, motivo per cui il loro impiego in ambito scolastico dovrebbe essere sottoposto a un attento controllo umano oppure evitato del tutto (Perkins et al, 2024). Affinché tali strumenti di intelligenza artificiale possano essere utilizzati nell'istruzione, i fornitori di sistemi di IA devono fornire spiegazioni sul modello e, in particolare, sulle sue limitazioni, basandosi su test approfonditi condotti su un set di dati appropriato che rifletta le caratteristiche degli utenti finali potenziali (in questo caso gli studenti).

Promuovere l'alfabetizzazione all'intelligenza artificiale tra insegnanti e studenti è inoltre fondamentale per comprendere questi strumenti e utilizzarli in modo responsabile. È importante che gli educatori siano consapevoli dei limiti dell'IA e adottino nella valutazione un approccio centrato sulla persona, poiché gli strumenti di IA dovrebbero supportare, e non sostituire, il giudizio umano. Gli insegnanti dovrebbero valutare il lavoro degli studenti in modo olistico, tenendo conto delle capacità individuali di ciascuno e fornendo un feedback personalizzato. Le politiche future dovrebbero tutelare gli approcci centrati sulla persona – inclusi metodi di valutazione alternativi, meccanismi di appello e una comunicazione chiara – per garantire l'equità. Maggiori informazioni sul tema della valutazione con l'utilizzo di strumenti di intelligenza artificiale sono disponibili nella relazione sull'IA redatta dalla prima squadra EDEH dedicata all'IA nell'istruzione.

Aspetti giuridici

Questo sistema di rilevamento dell'IA sarebbe classificato come ad alto rischio ai sensi dell'AI Act, in quanto incide direttamente sul percorso scolastico, sulle opportunità future e sul benessere emotivo di Emil (articolo 6, paragrafo 2, in combinato disposto con l'allegato III, paragrafo 3, lettera b)). Di conseguenza, il responsabile decisionale all'interno dell'istituto scolastico, ad esempio il dirigente scolastico, è tenuto a svolgere una valutazione d'impatto sui rischi legati ai diritti fondamentali (FRIA).

Questo passaggio garantisce che i rischi legati a distorsioni, equità e trasparenza siano identificati e mitigati per proteggere i diritti fondamentali degli studenti e assicurare pari opportunità (articolo 27, dell'AI Act). In conformità con i principi di trasparenza (articolo 13 dell'AI Act), gli sviluppatori dello strumento devono fornire al dirigente scolastico chiare istruzioni per l'uso e includere informazioni sulle limitazioni e sui rischi del sistema. Lo sviluppatore deve inoltre spiegare come interpretare i risultati, descrivere eventuali modifiche apportate al sistema che sono state predeterminate dal fornitore e includere misure di manutenzione, nonché descrivere come raccogliere, conservare e interpretare i dati di input (articolo 13, dell'AI Act). Inoltre, per ottemperare agli obblighi relativi alla sorveglianza umana (articolo 14, dell'AI Act), è indispensabile che il dirigente scolastico e gli altri utenti, come gli insegnanti, ricevano una formazione adeguata sui sistemi di IA, per comprenderne il funzionamento e, se necessario, annullare gli output automatici del sistema. Siffatto meccanismo di sorveglianza è fondamentale per tutelare gli studenti da decisioni potenzialmente errate o ingiuste. Pertanto, in questo caso, il preside della scuola dovrà adottare misure protettive tecniche e organizzative per garantire che il sistema sia utilizzato per lo scopo previsto e implementato in modo sicuro ed equo (articolo 26, dell'AI Act).

Ai sensi del GDPR, l'uso da parte della scuola di uno strumento di rilevamento basato sull'intelligenza artificiale per valutare il lavoro di Emil solleva notevoli preoccupazioni. L'articolo 22 del GDPR vieta le decisioni basate esclusivamente sul trattamento automatizzato se queste incidono in modo significativo sulle persone fisiche, come nel caso in esame, poiché Emil non ha superato una prova importante. Dal momento che la decisione si è basata in larga misura sui risultati prodotti dall'intelligenza artificiale, senza che fosse prevista alcuna forma di supervisione umana, i diritti di Emil sono stati violati. Gli obblighi di trasparenza (articoli 12-14 del GDPR) impongono alla scuola di informare gli studenti sull'uso degli strumenti di IA, sulla loro logica, sul loro impatto e sul ruolo decisionale nel contesto accademico. Emil ha il diritto di contestare la decisione, chiedere una spiegazione e richiedere la revisione umana dei risultati. Oltre a tali inottemperanze procedurali, la scuola deve dimostrare responsabilità ai sensi del GDPR svolgendo una valutazione d'impatto sulla protezione dei dati (DPIA), per garantire il rispetto dei principi di equità, trasparenza e non discriminazione stabiliti dal GDPR. Per porre rimedio al problema, la scuola dovrebbe riesaminare manualmente il caso di Emil, divulgare le proprie politiche in materia di IA e garantire che gli strumenti utilizzati siano equi e affidabili per tutti gli studenti.

Aspetti tecnici

Se la tecnica di IA alla base di questo strumento fosse una "grey-box" o una "white-box" (vedi sezione 1.3.), la spiegazione del modello sarebbe facile da comprendere. Tuttavia, gli attuali strumenti di rilevamento si basano sull'IA generativa (deep learning), che rende il loro ragionamento significativamente più complesso e meno trasparente. Lo scenario in cui il progetto di ricerca di Emil in storia ottiene una valutazione negativa, dopo che uno strumento di rilevamento dell'IA ha segnalato il suo lavoro come generato dall'IA, solleva un problema di classificazione. I modelli di classificazione prevedono etichette di classificazione specifiche e, nel contesto del caso d'uso, esistono due possibili classificazioni (classificazione binaria) con quattro diversi risultati possibili:

		Condizione prevista	
		Positiva (PP)	**Negativa (PN)**
Condizione effettiva	**Positiva (AP)**	Il modello contrassegna **correttamente** il lavoro dello studente come generato dall'IA, configurando **un vero positivo)**	Il modello contrassegna **erroneamente** il lavoro dello studente come NON generato dall'IA, configurando un **falso negativo.**
	Negativa (AN)	Il modello contrassegna **erroneamente** il lavoro dello studente come generato dall'IA, configurando un **falso positivo (il caso di Emil).**	Il modello contrassegna **correttamente** il lavoro dello studente come NON generato dall'IA, configurando un **vero negativo**.

Tabella 7: Modelli di classificazione.

L'errore più grave è rappresentato dal falso positivo, che accusa erroneamente uno studente di aver copiato, come è accaduto a Emil. Per ridurre questo rischio, gli insegnanti devono ricevere dai fornitori informazioni chiare sull'accuratezza e sui limiti dello strumento, nonché indicazioni su come interpretare i segnali di allarme, in modo da poter considerare il sistema come un assistente e non come unica fonte di verità. Anche i falsi negativi, ovvero i casi in cui gli studenti che utilizzano strumenti di IA non vengono rilevati, pongono problemi significativi, ma talvolta sono considerati meno gravi.

I fornitori che sviluppano modelli di rilevamento dell'IA utilizzano metriche di performance quali accuratezza, precisione, richiamo e punteggio F1 per valutare i propri modelli. L'accuratezza misura la correttezza complessiva, il richiamo riflette la capacità del modello di identificare i casi effettivi di testo generato dall'IA e la precisione indica la capacità di evitare falsi positivi. Il punteggio F1 trova un equilibrio tra l'identificazione accurata del testo generato dall'IA (precisione) e l'assenza di segnalazioni errate di lavori scritti da esseri umani (richiamo).

Sebbene le metriche possano indicare un'elevata performance tecnica di un modello, esse non forniscono di per sé una comprensione significativa dell'impatto dei risultati ottenuti nel contesto educativo e risultano spesso poco familiari al personale docente. Gli educatori devono avere una comprensione più chiara di come tali metriche si traducano in applicazioni pratiche all'interno dell'ambiente scolastico. Inoltre, una previsione errata può compromettere la fiducia nel processo valutativo. Ad esempio, un valore di recall pari al 99% implica che 1 studente su 100 potrebbe essere erroneamente accusato di comportamenti scorretti, come il plagio, con potenziali conseguenze significative sia sul piano educativo sia su quello psicologico.

In tali circostanze, risulta fondamentale esplicitare il processo attraverso cui il modello giunge alle sue conclusioni, al fine di consentire agli insegnanti di giustificare le decisioni assunte dallo strumento e di rafforzare la fiducia degli studenti, chiarendo le basi delle valutazioni effettuate dall'intelligenza artificiale. Dal punto di vista dell'educatore, le spiegazioni basate sulla rilevanza delle caratteristiche testuali – come la struttura sintattica, la frequenza delle parole, e altri aspetti linguistici – risultano particolarmente utili quando sono espresse in modo chiaro e accessibile. Ciò consente agli insegnanti di interpretare in modo accurato le previsioni del modello e di comunicare efficacemente con gli studenti riguardo alle ragioni per cui un elaborato è stato segnalato o meno.

Sfide

La trasparenza costituisce una sfida qualora lo strumento di IA funzioni come una "scatola nera", rendendo difficile spiegarne la logica. Garantire una sorveglianza umana significativa richiede molte risorse, poiché è necessario disporre di personale qualificato per esaminare in modo equo i casi segnalati e per comprendere i dati di addestramento e il modello del sistema.

Raccomandazioni

Le scuole devono garantire trasparenza nell'uso degli strumenti di IA per la valutazione, comunicando in modo chiaro e diretto il ruolo di tali strumenti e le politiche associate. Al fine di promuovere l'equità, tali strumenti devono essere convalidati in termini di accuratezza e idoneità per diverse popolazioni di studenti, con la previsione della sorveglianza umana come parte integrante del processo decisionale. Prima dell'implementazione, le scuole devono condurre FRIA per i sistemi di IA ad alto rischio e DPIA per conformarsi ai principi di equità, trasparenza e non discriminazione nel trattamento dei dati personali ai sensi del GDPR. Gli educatori devono essere informati dai rispettivi responsabili delle decisioni in materia di istruzione (ad esempio i presidi) del fatto che questo tipo di sistema di intelligenza artificiale è ad alto rischio e non è affidabile. Si raccomanda quindi di evitare qualsiasi tipo di attività di apprendimento autonomo che possa essere svolta dagli studenti con IA generativa e di richiedere una supervisione antiplagio. Dopo l'implementazione, gli studenti dovrebbero avere la possibilità di contestare le decisioni automatizzate dell'IA, rendendosi pertanto necessario istituire un meccanismo di ricorso per salvaguardare i diritti degli studenti. Da un punto di vista tecnico, è fondamentale adottare un approccio interdisciplinare che contestualizzi la valutazione dei modelli di IA, integrando considerazioni tecniche, etiche ed educative. I fornitori di IA devono fornire spiegazioni interpretabili dei processi decisionali dei loro sistemi, utilizzando metodi post-hoc come la rilevanza delle caratteristiche (ad esempio SHAP o LIME). I fornitori dovrebbero garantire trasparenza riguardo alle caratteristiche valutate dal modello, offrendo formazione e documentazione che spieghino come tali caratteristiche siano correlate al testo generato dall'intelligenza artificiale. In ultima analisi, la responsabilità non ricade solo sulle scuole, ma anche sugli sviluppatori, che devono essere ritenuti responsabili di garantire che i loro sistemi siano trasparenti, equi e ben documentati, permettendone un uso responsabile nei contesti educativi.

Sistema di tutoraggio intelligente

Julia, un'alunna di terza elementare con una lieve dislessia, sta imparando l'inglese come seconda lingua con l'utilizzo di un nuovo libro di testo digitale basato sull'intelligenza artificiale, progettato per personalizzare l'apprendimento della matematica e dell'inglese. Questo sistema adatta i contenuti alle sue difficoltà e ai suoi punti di forza specifici. Il libro di testo evidenzia i termini chiave, offre formulazioni semplificate e fornisce icone visive per le parole complesse. In matematica, si adatta al ritmo più lento di Julia e previene la confusione tra numeri come 47 e 74.

Gli esercizi sono suddivisi in piccoli passaggi e Julia riceve feedback immediati ed esempi interattivi. In inglese, suggerimenti audio, traduzioni e un vocabolario semplificato la aiutano a comprendere il testo. Man mano che Julia progredisce, il libro di testo si adatta per proporle compiti più impegnativi, pur mantenendo gli strumenti di supporto. Il libro di testo digitale consente inoltre a Julia di monitorare i propri progressi attraverso report visivi che evidenziano i suoi punti di forza e le aree in cui può migliorare. Tracciando i modelli di apprendimento, l'IA può non solo supportare la crescita accademica di Julia, ma anche rafforzare la consapevolezza di sé come studente. I genitori e gli insegnanti di Julia possono accedere a una dashboard che fornisce informazioni dettagliate sul suo percorso di apprendimento, aiutandoli a comprendere il supporto che sta ricevendo Julia e i progressi compiuti.

Aspetti educativi

Per rendere questi strumenti più efficaci ed equi, sono essenziali la trasparenza e la supervisione umana. È auspicabile che gli educatori collaborino attivamente con i sistemi di IA, al fine di validare le raccomandazioni prodotte e offrire un feedback personalizzato agli studenti (Sağın et al, 2024). È altresì importante che gli educatori abbiano la possibilità di intervenire sulle raccomandazioni dell'IA e di adattarle a esigenze e contesti specifici. I framework inclusivi di intelligenza artificiale applicati in contesti educativi dovrebbero implementare approcci quali la co-creazione, al fine di assicurare che le tecnologie rispondano efficacemente alla pluralità dei bisogni di discenti ed educatori, che siano garantite adeguate salvaguardie etiche, e che siano promossi i valori di inclusione ed equità. Il design partecipativo pone l'accento sul coinvolgimento di diversi stakeholder — quali discenti, educatori e genitori — affinché gli strumenti di IA rispondano a esigenze culturali, linguistiche e pedagogiche differenti. Principi etici come quelli enunciati nella Raccomandazione dell'UNESCO sull'etica dell'intelligenza artificiale (2021) e gli Orientamenti etici per gli educatori sull'uso dell'intelligenza artificiale (IA) e dei dati nell'insegnamento e nell'apprendimento dell'UE (2022), sottolineano l'importanza dell'inclusività, della non discriminazione e della riduzione delle disuguaglianze nei sistemi di IA applicati in ambito educativo. Anche politiche chiare in materia di protezione dei dati sono fondamentali per salvaguardare la privacy degli studenti.

Aspetti legali

Ai sensi dell'AI Act, questo scenario sarebbe considerato un'applicazione di intelligenza artificiale ad alto rischio (articolo 6, paragrafo 2, in combinato disposto con l'allegato III, paragrafo 3, lettera b)), a causa del ruolo del sistema di IA nelle decisioni in ambito educativo, della personalizzazione basata sui dati e del potenziale impatto sui risultati di apprendimento di Julia. È pertanto necessario condurre una FRIA, poiché il sistema di IA utilizzato comporta la profilazione ad alto rischio di una bambina vulnerabile, con un impatto su diritti fondamentali quali la privacy, l'uguaglianza e l'istruzione. Il fornitore è tenuto a progettare il sistema di IA in modo tale da consentire a Julia, agli educatori e ai suoi genitori, di comprendere come vengono prese le decisioni automatizzate, rendendoli in grado di valutare le motivazioni alla base degli adattamenti e delle raccomandazioni personalizzati (articolo 13, paragrafo 1, dell' AI Act).

Inoltre, gli sviluppatori devono progettare il sistema di IA in modo tale da fornire istruzioni per l'uso concise, complete, corrette e chiare, affinché risultino pertinenti, accessibili e comprensibili per gli operatori che utilizzano il sistema in ambito scolastico (articolo 13, paragrafo 2, dell'AI Act). Anche la governance dei dati (articolo 10; considerando 66, 67 e 69 dell'AI Act) è fondamentale, al fine di garantire che i dati utilizzati siano pertinenti, rappresentativi, privi di errori e completi. Per evitare risultati distorti, è necessario effettuare test regolari, utilizzando set di dati diversificati che riflettano la diversità demografica e linguistica degli studenti. La sorveglianza umana (articolo 14, dell'AI Act) continua ad essere fondamentale e gli insegnanti di Julia dovrebbero mantenere la possibilità di intervenire o di ignorare le raccomandazioni formulate dall'IA. Ne discende che gli sviluppatori dovrebbero progettare il sistema di IA in modo tale da permettere agli operatori (in questo caso la scuola e gli insegnanti) di esercitare una supervisione efficace, anche mediante adeguati strumenti di interfaccia uomo-macchina. Data la natura sensibile delle informazioni di Julia, comprese le sue difficoltà di apprendimento e i suoi progressi, sono applicabili diverse disposizioni del GDPR che operano parallelamente all'AI Act (ad esempio, l'articolo 10, considerando 69). Ai sensi di tali disposizioni dovrebbero essere raccolti solo i dati essenziali di Julia, come la sua capacità di lettura, evitando la raccolta di dati non necessari (principio di minimizzazione dei dati).

È imperativo che i dati raccolti vengano impiegati esclusivamente per lo scopo originario e non riutilizzati in attività estranee, quali il marketing, in conformità con il principio della limitazione delle finalità. Qualora la scuola intendesse impiegare tali dati diversamente, dovrà ottenere il consenso esplicito dei genitori o dei tutori di Julia. Occorre inoltre prevedere meccanismi trasparenti e facilmente accessibili che descrivano l'uso dei dati e ne consentano l'immediata revoca del consenso.

Lo strumento deve garantire la tutela della privacy mediante l'impiego di tecniche come l'anonimizzazione, al fine di salvaguardare l'identità della studentessa. Julia e i suoi genitori devono poter esercitare i diritti di accesso, rettifica o cancellazione dei dati. In aggiunta, l'analisi d'impatto sulla protezione dei dati (DPIA) deve essere svolta per valutare i rischi e individuare misure correttive, assicurando trasparenza e responsabilità nel trattamento. Quando lo strumento opera su una piattaforma online, trova applicazione anche il regolamento sui servizi digitali (DSA). Il DSA richiede trasparenza negli algoritmi, in particolare per quanto riguarda la fornitura di contenuti personalizzati, come esercizi su misura o relazioni sui progressi compiuti. Inoltre, la piattaforma online deve proteggere i minori da contenuti dannosi (articolo 28, del DSA). Se la piattaforma di apprendimento utilizza meccanismi come impostazioni predefinite per il consenso al trattamento esteso dei dati, opzioni di adesione preselezionate o notifiche confuse che spingono ad aggiornamenti non necessari, tali meccanismi potrebbero essere considerati dark pattern (articolo 23 bis, paragrafo 1, considerando 51, lettera b) del DSA) , e comportare l'adozione di misure di esecuzione.

Aspetti tecnici

Il libro di testo digitale di Julia basato sull'intelligenza artificiale enfatizza il ruolo fondamentale dell'IA spiegabile nel garantire trasparenza, fiducia e responsabilità lungo tutto il processo educativo. Da un lato, il design e i contenuti della dashboard sono determinanti per decidere cosa spiegare a Julia. Tale dashboard dovrebbe fornire visualizzazioni chiare e complete, volte a illustrare i progressi di Julia e a descrivere le logiche sottostanti agli interventi adattivi. Dall'altro lato, la protezione dei dati personali è imprescindibile, soprattutto poiché informazioni sensibili su eventuali difficoltà, come la dislessia, devono essere trattate secondo le normative vigenti, tra le quali il DATA Act dell'UE.

In questo caso d'uso, la tutela dei dati sanitari di Julia richiede non solo una spiegazione delle predizioni del modello, ma anche un livello più elevato di trasparenza, esteso all'hosting del modello e all'intero processo decisionale. Si assume che il sistema di tutoring intelligente (ITS) includa un modello in grado di rilevare la condizione di Julia e generare raccomandazioni personalizzate sulla base di tale condizione. L'adozione della XAI è pertanto fondamentale, poiché il modello potrebbe confondere la condizione di Julia con altre, contribuendo a interpretazioni errate — ad esempio, l'attribuzione erronea dei suoi pattern di apprendimento alla dislessia o ad altre condizioni. Per prevenire bias e giustificare adeguate adattazioni, sono necessarie spiegazioni globali, quali la rilevanza delle caratteristiche, al fine di comprendere come il modello associa i pattern di interazione dell'alunno al contenuto didattico. Allo stesso tempo, Julia e i suoi educatori potrebbero richiedere una spiegazione locale da parte dei fornitori del sistema riguardo alle previsioni ricevute, per assicurarsi che il modello stia effettivamente rispondendo alla sua dislessia e non stia generando una previsione basata su un'ipotesi errata. Attraverso l'integrazione di approcci Human-in-the-Loop (HITL), il sistema consente a educatori e genitori di intervenire, garantendo che i bisogni di apprendimento di Julia vengano soddisfa.

Sfide

Le principali sfide consistono nel garantire che i sistemi siano conformi alle disposizioni vigenti, le quali richiedono trasparenza, spiegabilità, rispetto della privacy, supervisione umana e tutela dei diritti di Julia. Le priorità fondamentali includono una solida governance dei dati, la scongiura di eventuali di bias, la predisposizione di meccanismi chiari per la prestazione del consenso e la protezione da contenuti dannosi o pratiche manipolative.

Raccomandazioni

Gli istituti di istruzione devono garantire un uso trasparente degli strumenti di IA nell'insegnamento e nell'apprendimento, comunicando chiaramente agli educatori, ai genitori e agli studenti il ruolo di tali strumenti e le politiche associate. Per promuovere l'equità, questi strumenti devono essere validati in termini di accuratezza e adeguatezza rispetto a diverse popolazioni di discenti, con la supervisione umana come parte integrante del processo decisionale. Le istituzioni devono condurre valutazioni di impatto sui diritti fondamentali (FRIA) dei sistemi di intelligenza artificiale ad alto rischio e valutare l'impatto di tali sistemi sulla protezione dati (DPIA), pper garantire l'osservanza dei principi del GDPR relativi a correttezza, trasparenza e non discriminazione nel trattamento dei dati personali. Gli educatori devono essere informati dalle autorità scolastiche competenti (ad esempio, il dirigente) del fatto che questo tipo di sistema di intelligenza artificiale è considerato ad alto rischio e deve essere valutato prima di essere utilizzato con gli studenti. Dal punto di vista dell'educatore, è essenziale che l'ITS venga valutato anche in termini di efficacia e affidabilità rispetto agli approcci didattici e al design dell'apprendimento su cui è stato addestrato, nonché rispetto al supporto effettivo che fornisce a tutti gli studenti, tenendo conto dei loro bisogni educativi speciali e del loro stile di apprendimento. È fondamentale che gli sviluppatori e i fornitori di ITS garantiscano una sorveglianza umana integrata fin dalla progettazione, in modo da permettere agli educatori di intervenire sulle decisioni dell'ITS e assegnare "manualmente" i compiti o apportare modifiche ai percorsi di apprendimento.

Come anticipato, gli educatori dovrebbero inoltre essere dotati di strumenti per supervisionare, intervenire o ignorare le raccomandazioni dell'IA, al fine di garantire l'uso responsabile degli strumenti e impedire l'affidamento esclusivo ai sistemi automatizzati. I sistemi dovrebbero essere progettati per fornire a Julia, ai suoi genitori e agli insegnanti spiegazioni chiare e accessibili delle decisioni automatizzate, insieme a istruzioni di facile utilizzo per la scuola. Da una prospettiva tecnica, un approccio interdisciplinare è fondamentale per contestualizzare la valutazione dei modelli di intelligenza artificiale, integrando considerazioni tecniche, etiche ed educative. Pertanto, è necessario che i fornitori di IA garantiscano spiegazioni interpretabili dei processi decisionali delle loro soluzioni, preferibilmente attraverso metodi post-hoc come quelli basati sulla rilevanza delle caratteristiche, quali **SHAP** o **LIME**. I fornitori dovrebbero altresì garantire la trasparenza rispetto alle caratteristiche valutate dal modello, offrendo la formazione e la documentazione necessarie a spiegare come tali caratteristiche siano correlate ai testi generati dall'intelligenza artificiale. In definitiva, la responsabilità non ricade solo sulle istituzioni utilizzatrici, ma si estende anche agli sviluppatori, che devono essere ritenuti responsabili nel garantire che i loro sistemi siano trasparenti, equi e ben documentati, per permetterne un uso responsabile nei contesti educativi.

Valutazione automatizzata

Un'università adotta un sistema di valutazione automatizzato basato sull'intelligenza artificiale che utilizza un indice IA per valutare i saggi degli studenti. Sebbene il sistema valuti rapidamente le consegne in base alla struttura e alla correttezza lessicale, i docenti notano che gli studenti provenienti da contesti linguistici diversi ottengono costantemente punteggi più bassi. Questa tendenza solleva preoccupazioni circa la presenza di bias sistemici, poiché l'algoritmo sembra svantaggiare gli studenti che utilizzano stili linguistici diversi o di lingua madre straniera. In risposta alle disparità percepite, gli studenti, supportati dai genitori, esprimono la loro frustrazione, mettendo in discussione l'equità di valutazioni operate da un sistema opaco e chiedono quindi maggiore trasparenza, con la garanzia che il loro potenziale accademico non venga compromesso da bias nascosti.

Aspetti educativi

Questo caso evidenzia le sfide e le implicazioni etiche dell'implementazione di sistemi di valutazione basati sull'intelligenza artificiale utilizzati nel settore educativo. L'uso di tali sistemi per valutare i saggi ha portato a conseguenze indesiderate, con la conseguenza che studenti provenienti da contesti linguistici, culturali ed economici diversi hanno ottenuto punteggi costantemente più bassi. Questa tendenza solleva notevoli preoccupazioni in merito alle distorsioni sistemiche, alla trasparenza e all'inclusività degli strumenti di valutazione basati sull'IA. Una questione fondamentale è l'apparente parzialità dell'algoritmo di valutazione. Gli studenti che utilizzano stili linguistici diversi o che sono di lingua madre straniera potrebbero non conformarsi ai modelli che l'IA associa a una scrittura di qualità superiore (Wang, 2024). Ciò non solo influisce sui loro voti, ma potenzialmente mina anche la loro fiducia e di conseguenza il loro progresso accademico.

Le distorsioni appena descritte evidenziano il rischio di affidarsi a sistemi di IA che non sono in grado di adattarsi alla diversità delle espressioni linguistiche e culturali, come parte integrante di un ambiente accademico globalmente interconnesso. La mancanza di trasparenza del sistema aggrava tali preoccupazioni. Gli studenti e i genitori non sanno come vengono assegnati i voti, il che alimenta la sfiducia nel sistema. Senza spiegazioni chiare sul funzionamento dell'IA e sui criteri che essa privilegia, gli studenti non possono contestare o mettere in discussione i voti ricevuti in modo efficace. Quindi la trasparenza è essenziale non solo per garantire l'equità, ma anche per costruire la fiducia negli strumenti automatizzati. Un sistema XAI potrebbe fornire un valido feedback sul motivo per cui è stato assegnato un determinato punteggio, aiutando gli studenti a comprendere le loro prestazioni e a migliorare.

Per affrontare queste sfide, gli istituti di istruzione devono adottare un approccio antropocentrico per la valutazione con strumenti di IA (Topali et al, 2024). Gli educatori devono quindi assumere un ruolo attivo nella validazione degli algoritmi di valutazione e dei voti generati dall'intelligenza artificiale, assicurandosi che eventuali bias vengano individuati e corretti. Infatti, l'intelligenza artificiale dovrebbe fungere da strumento complementare, non da sostituto del giudizio umano, con valutazioni finali che integrino un feedback qualitativo in grado di valorizzare i diversi contributi linguistici e culturali. In fine, gli studenti devono essere informati sul funzionamento del sistema e avere la possibilità di partecipare a valutazioni alternative o a procedure di ricorso in caso di discrepanze.

Aspetti legali

Il sistema del caso di specie è ad alto rischio ai sensi della legge sull'IA, in particolare dell'allegato III, che inserisce in questa classe i sistemi che valutano i risultati dell'apprendimento in contesti educativi. È probabile che sia necessaria una FRIA, poiché il sistema di valutazione basato sull'IA solleva preoccupazioni in termini di distorsioni, equità e trasparenza, con potenziali ripercussioni sui diritti fondamentali e sulle pari opportunità degli studenti. La mancanza di spiegazioni chiare sulle decisioni relative alla valutazione viola gli obblighi di trasparenza (articolo 13 dell'AI Act), che impongono ai fornitori di sviluppare sistemi di IA che consentano di accedere alle informazioni sul funzionamento del sistema, nonché sui suoi limiti e rischi. Inoltre, gli studenti e i genitori hanno diritto a una spiegazione ai sensi dell'articolo 22 del GDPR. Il sistema manca di una sorveglianza significativa (articolo 14 dell' AI Act). Inoltre, lo svantaggio arrecato a chi non è di madrelingua italiana rivela la presenza di un bias sistemico, violando i requisiti di governance dei dati (articolo 10 dell' AI Act), che richiedono dati di addestramento rappresentativi e non discriminatori.

È possibile che il sistema di valutazione operi tramite una piattaforma online, soprattutto se l'università utilizza un sistema di gestione dell'apprendimento digitale (LMS) più ampio. In tal caso, si applicherebbero le disposizioni del DSA in materia di trasparenza algoritmica e diritti degli utenti (articolo 24, del DSA). Il regolamento sulla governance dei dati (DGA) è altresì rilevante se il sistema di valutazione utilizza set di dati condivisi o collabora con intermediari di dati affidabili per gestire i dati per l'addestramento del modello di IA. Anche il Data Act (DA), che si concentra sulla condivisione sicura dei dati e sull'interoperabilità, è applicabile al caso di specie se il sistema di valutazione condivide dati con altri sistemi (ad esempio, piattaforme di reportistica o banche dati istituzionali), oppure se si basa su set di dati esterni per i suoi processi di addestramento e valutazione.

Aspetti tecnici

La valutazione automatica di testi liberi, come quelli contenuti negli elaborati degli studenti, è un compito impegnativo per i sistemi e per gli algoritmi di intelligenza artificiale. Ciò è dovuto al fatto che tale forma di valutazione riguarda una serie complessa di criteri di accettazione, tra cui l'autenticità, il contesto, lo stile lessicale e le capacità linguistiche. Un sistema di valutazione basato sull'intelligenza artificiale deve funzionare a livello semantico piuttosto che a livello di singole parole. In altri termini, deve essere in grado di riconoscere il significato delle frasi e non limitarsi ad analizzare il vocabolario utilizzato.

Gli educatori e i responsabili del settore educativo che intendono adottare sistemi di valutazione automatica richiedono agli sviluppatori una descrizione chiara e trasparente delle modalità attraverso cui tali sistemi generano le valutazioni, considerate indicative della "correttezza" o dell'"autenticità" di un elaborato. In questo contesto, gli approcci e le tecniche di XAI risultano particolarmente efficaci, in quanto permettono di evidenziare le caratteristiche testuali che influenzano le previsioni del modello, fornendo così informazioni utili sia sulla rilevanza delle variabili considerate sia sul funzionamento complessivo del sistema..

Ad esempio, gli approcci basati sulla rilevanza delle caratteristiche nellaXAI, come l'importanza delle caratteristiche di permutazione (PFI) e i grafici di dipendenza parziale (PDP), hanno il potenziale di chiarire agli educatori e ai responsabili dell'istruzione se esistono caratteristiche dominanti sulle quali il sistema si basa per determinare il voto di un saggio. La PFI misura l'impatto delle singole caratteristiche (ad esempio, la qualità grammaticale, la struttura, la ricchezza del vocabolario) sulle previsioni del modello permutando (randomizzando) i valori di una singola caratteristica e osservando il cambiamento risultante nelle prestazioni del modello. In tal modo si riesce a capire se alcune caratteristiche correlate al background dello studente stanno influenzando l'output del modello, creando previsioni distorte. In questo caso d'uso, gli educatori dovrebbero poter richiedere ai fornitori spiegazioni PFI relative a caratteristiche specifiche, come la diversità lessicale o la complessità grammaticale, considerando che queste sono tra le caratteristiche che riflettono la diversità del background degli studenti. Sulla stessa linea, i PDP possono visualizzare la relazione tra le caratteristiche e il risultato previsto, ovvero il punteggio del saggio. Il loro vantaggio è che mostrano il comportamento del modello su una gamma di valori della caratteristica. Considerando la caratteristica della complessità grammaticale, ad esempio, un PDP può mostrare se l'aumento della complessità grammaticale nel saggio comporta sempre un aumento del punteggio finale. Questo dato potrebbe indicare un bias del sistema di valutazione.

Tuttavia, la questione di come raggiungere una comprensione adeguata è complessa, soprattutto perché esiste una componente soggettiva nel definire cosa sia o meno una buona spiegazione. Una descrizione comprensibile può, in ultima analisi, essere soggettiva. Per rispondere a questa domanda sono necessarie definizioni chiare e un allineamento tra tutte le parti interessate, come sottolineato più volte nel presente rapporto.

È inoltre importante notare che la distorsione del modello può derivare dall'intero processo di addestramento e hosting. I dati di addestramento raccolti principalmente da un solo gruppo di studenti determineranno una distorsione nelle previsioni del sistema, favorendo i saggi di questo gruppo. L'identificazione dell'eventuale presenza di distorsioni nei sistemi di intelligenza artificiale, l'analisi delle loro cause e la definizione di strategie di mitigazione costituiscono una responsabilità primaria dei fornitori di tali sistemi, i quali, in quanto conoscitori del processo di raccolta e trattamento dei dati, sono tenuti a garantirne la trasparenza nei confronti degli stakeholder del settore educativo. Tuttavia, gli stakeholder possono offrire un contributo significativo, ad esempio proponendo criteri per la valutazione e la mitigazione delle distorsioni. Ne consegue un approccio necessariamente interdisciplinare, in cui educatori e sviluppatori collaborano attivamente per identificare e ridurre i bias presenti nei sistemi di valutazione automatica.

Sfide

Una sfida importante è quella di garantire che i dati di addestramento siano diversificati e rappresentativi, per evitare il perpetuarsi di discriminazioni sistematiche. Un'ulteriore sfida è rappresentata dal fatto che il sistema funziona come una "scatola nera" e quindi gli studenti, i genitori e i docenti non sono in grado di comprendere o contestualizzare le decisioni, a scapito dei principi di fiducia e di responsabilità. Un'altra sfida è quella di trovare un equilibrio tra l'automazione dell'IA e la funzione di sorveglianza umana senza generare inefficienze operative.

Raccomandazioni

L'autorità educativa dovrebbe sospendere l'uso del sistema di valutazione e coinvolgere le parti interessate, come il personale, i genitori e gli studenti, in un processo di revisione. Il responsabile dell'istruzione dovrebbe chiedere spiegazioni a livello globale per chiarire come il sistema valuti i compiti e spiegazioni a livello locale per aiutare a comunicare agli studenti perché sono stati assegnati determinati voti e quali caratteristiche hanno influenzato il risultato. Gli educatori dovrebbero ricevere una formazione sull'uso del sistema per garantire una solida sorveglianza umana. Inoltre, l'università dovrebbe stabilire misure di tutela per garantire l'allineamento del sistema di IA con gli standard etici e legali. Anche in questo caso, i fornitori di sistemi di intelligenza artificiale devono offrire spiegazioni comprensibili dei processi decisionali dei loro sistemi, garantendo trasparenza rispetto alle caratteristiche valutate dal modello e offrendo formazione e documentazione che spieghino come tali caratteristiche siano correlate alla valutazione. Per conformarsi alle disposizioni legali specifiche sarebbe necessario un approccio collaborativo che coinvolga tutte le parti interessate. È importante che gli istituti di istruzione conducano valutazioni d'impatto complete, ad esempio di tipo FRIA (AI Act, per i sistemi ad alto rischio) e DPIA (GDPR, per il trattamento dei dati personali), al fine di valutare i rischi e garantire la conformità prima che questo tipo di sistemi sia messo in funzione. In termini di governance dei dati, gli sviluppatori devono raccogliere dati da fonti diverse per garantire che siano pertinenti e rappresentativi e altresì garantire l'integrità dei dati per mitigare le distorsioni sistemiche (articolo 10,dell'AI Act).

Inoltre, ai sensi dell'AI Act, gli sviluppatori devono predisporre canali accessibili (ad esempio dashboard) per fornire a educatori e studenti spiegazioni chiare e concise delle decisioni relative alla valutazione, nonché istruzioni d'uso dettagliate (articolo 13). Gli sviluppatori devono progettare sistemi con interfacce uomo-macchina che consentano agli educatori di intervenire o di ignorare le decisioni, e i responsabili dell'istruzione dovrebbero monitorare attivamente il sistema e utilizzare il proprio giudizio per garantirne l'equità (articolo 14). È altrettanto importante che gli istituti di istruzione rispettino i principi del GDPR, compresa la minimizzazione dei dati, quando trattano dati personali (articolo 5, GDPR). Gli istituti dovrebbero inoltre rendere anonimi i dati per proteggere i dati personali e limitare la raccolta solo a quanto strettamente necessario per la finalità prevista. Inoltre, gli studenti e i genitori dovrebbero svolgere un ruolo attivo nella gestione del consenso all'utilizzo dei dati. Dal canto loro, gli sviluppatori devono prevenire i modelli oscuri (articolo 23 bis, paragrafo 1, e considerando 51, lettera b), del DSA) e garantire la trasparenza algoritmica nella fornitura di contenuti personalizzati (articolo 28, del DSA). In fine, le autorità di regolamentazione dovrebbero vigilare sul rispetto di tali disposizioni, mentre gli istituti di istruzione svolgono un ruolo nel dare priorità alle piattaforme per la protezione degli studenti, in particolare dei minori, da eventuali danni.

2.4 Come implementare l'IA in modo responsabile

Al fine di fornire raccomandazioni semplici e pratiche alla comunità educante in merito all'implementazione responsabile dell'IA nell'istruzione ai sensi delle normative applicabili, la seguente tabella riassume le idee chiave emerse dai precedenti casi d'uso

Non affidarsi esclusivamente allo strumento	Comunicare chiaramente con gli studenti	Formare gli educatori e il personale	Predisporre misure di tutela	Valutare regolarmente lo strumento
Gli strumenti di IA dovrebbero fungere da supporto senza sostituire il giudizio dell'educatore, che dovrà sempre rivedere manualmente i lavori segnalati dal sistema prima di prendere una decisione	Spiegare agli studenti come funziona lo strumento; cosa fa, come lo fa e cosa non fa Essere trasparenti circa il ruolo dell'IA nella valutazione o nella revisione dei lavori	Assicurarsi che il personale capisca le finalità e le limitazioni dello strumento e sappia interpretarne i risultati Fornire indicazioni sulla gestione dei falsi positivi	Prevedere una procedura di ricorso che permetta agli studenti di contestare i risultati ingiusti Documentare le decisioni e garantire che le revisioni siano eque e imparziali	Verificare che lo strumento stia segnalando I lavori in modo equo per tutti i gruppi di studenti Chiedere al fornitore metriche di accuratezza e caratteristiche di spiegabilità

Tabella 8: Idee chiave per un'implementazione responsabile dell'IA nell'istruzione.

2.5 Aree chiave e punti di attenzione in merito all'implementazione dell'IA

Le spiegazioni fornite dai sistemi di intelligenza artificiale rivestono un ruolo cruciale per una efficace implementazione dell'IA in ambito educativo, in quanto consentono a docenti e studenti di fornire un riscontro informato sull'affidabilità e sull'utilità degli strumenti adottati. Tale feedback sarà in parte il risultato di un processo condiviso di interpretazione e comprensione del funzionamento dei sistemi in uso. Al fine di mitigare potenziali rischi, è indispensabile garantire una comunicazione trasparente che permetta agli attori coinvolti di esprimere eventuali preoccupazioni o criticità. Inoltre, è fondamentale evitare l'introduzione di sistemi di IA la cui complessità possa ostacolare l'attività degli educatori o distoglierli dalle loro principali funzioni didattiche.

Ci sono alcuni pro e contro da considerare. Se l'intelligenza artificiale permette di offrire un supporto personalizzato agli studenti, il lavoro dei docenti potrebbe diventare più efficace ed efficiente. Tuttavia, esiste il rischio che il tempo a disposizione venga assorbito da strumenti complessi da usare o dalla necessità di spiegarne il funzionamento agli studenti. Le spiegazioni dell'intelligenza artificiale sulle quali gli educatori vengono formati e che forniranno agli studenti non devono riguardare prodotti specifici (poiché ciò comporterebbe una forte dipendenza da determinati sistemi), bensì devono essere orientate alla comprensione del funzionamento generale dei sistemi. Gli insegnanti avranno bisogno di formazione e supporto, e questo deve essere un elemento da considerare nell'adozione delle tecnologie da parte della scuola. Se vi sarà un'ampia diffusione di sistemi di IA nell'istruzione, la formazione dei docenti – in servizio o tirocinanti -, dovrà includere tali argomenti. Un problema della XAI è che potrebbe indurre le persone a credere che le spiegazioni possano risolvere anche problemi che in realtà non sono in grado di affrontare, o che le spiegazioni siano più importanti dei problemi stessi da risolvere**.** Devono esserci casi d'uso chiari, basati sulla reale esperienza delle scuole, in grado di dimostrare che i risultati educativi migliorano quando l'IA entra in classe[16]

La tabella seguente illustra le aree chiave che gli istituti di istruzione, gli sviluppatori e i responsabili politici dovrebbero prendere in considerazione quando adottano sistemi di IA. Queste categorie riflettono le priorità condivise a diversi livelli del sistema educativo; considerate congiuntamente, esse offrono **una tabella di marcia realistica ed equilibrata per un'adozione responsabile dell'IA nell'istruzione**.

16 Per ulteriori raccomandazioni politiche si vedano gli esiti del workshop EDEH sulla XAI nell'istruzione, tenutosi il 17-18 ottobre 2024 a Bruxelles.

Voce	Categoria
Stabilire chiari meccanismi di feedback: • **Sviluppare canali di feedback per consentire alle parti interessate di scambiare osservazioni sulle prestazioni dei sistemi di IA.** • **Garantire che le spiegazioni siano co-costruite dalle parti interessate per migliorare la comprensione reciproca delle funzionalità del sistema.**	Meccanismi di feedback
Progettare sistemi di facile utilizzo per gli educatori: • **Evitare sistemi che impongono agli insegnanti requisiti operativi complessi, sottraendo tempo e risorse alle funzioni didattiche principali.** • **Adottare sistemi autoesplicativi, che non richiedano agli insegnanti di dedicare un tempo eccessivo per spiegare agli studenti come funzionano.**	Sistemi di facile utilizzo per gli educatori
Fornire una formazione generalizzabile agli educatori • **Formare gli educatori affinché siano in grado di spiegare il funzionamento scientifico dei sistemi di IA ed evitare una formazione specifica sui prodotti per prevenire dipendenze da prodotti specifici (vendor lock-in).** • **Incorporare la formazione nei programmi di sviluppo professionale per i tirocinanti e per i docenti in servizio.**	Formazione
Concentrarsi sui risultati educativi • **Valutare e adottare sistemi di IA basati su casi d'uso chiari e dimostrabili, in linea con gli obiettivi di miglioramento dei risultati didattici.** • **Bilanciare l'importanza delle spiegazioni con i problemi reali che i sistemi di IA sono progettati per risolvere.**	Risultati educativi
Sostenere l'adozione efficace delle tecnologie • **Includere i requisiti di formazione e supporto nel processo decisionale per l'adozione dei sistemi di IA nelle scuole.** • **Garantire la disponibilità continua di risorse per lo sviluppo professionale degli educatori, affinché possano adattarsi all'introduzione e all'utilizzo dell'intelligenza artificiale.**	Adozione delle tecnologie
Mitigare i rischi e affrontare le tensioni • **Creare canali per segnalare preoccupazioni riguardo ai sistemi di IA, assicurandosi che non sovraccarichino gli insegnanti né compromettano la loro efficacia didattica.** • **Rivedere e adattare regolarmente le implementazioni dell'IA per garantire che supportino, e non ostacolino, un insegnamento e un apprendimento efficaci.** **Promuovere casi d'uso chiari e trasparenti** • **Mostrare esempi concreti di implementazioni riuscite dei sistemi di IA per costruire fiducia tra le parti interessate.** • **Utilizzare programmi pilota per dimostrare miglioramenti misurabili nei processi e nei risultati educativi.**	Mitigazione dei rischi

Tabella 9 Check list ACE: Adempimenti per l'utilizzo dell'IA nell'istruzione.

3. La XAI nell'istruzione dal punto di vista dei vari stakeholder

3.1 Contesto

Dopo aver analizzato le questioni legali relative alla XAI nel capitolo precedente, e prima di presentare, nel capitolo 4, le competenze richieste agli educatori per integrare la XAI con sicurezza nella scuola, è opportuno illustrare cosa significhi includere la spiegabilità nell'istruzione da una prospettiva pratica. A tal fine, il presente capitolo riprende il tema della *comprensibilità* della XAI, già affrontato nella sezione 1.3., ovvero la necessità che i diversi utenti finali comprendano correttamente le spiegazioni fornite, al fine di rafforzare la loro fiducia nei sistemi di IA. Gli stakeholder definiti nella sezione 1.5 sono di fondamentale importanza in questo capitolo, poiché le loro prospettive individuali saranno il focus dell'analisi.

L'impatto della XAI viene esplorato attraverso due strumenti educativi basati sull'intelligenza artificiale, in particolare **i sistemi di tutoraggio intelligente (ITS) e i generatori di piani didattici basati sull'intelligenza artificiale (LPG)** che, essendo progettati per migliorare l'apprendimento personalizzato, si adattano alle esigenze individuali degli studenti e supportano gli educatori nella creazione di contenuti didattici su misura. Tali strumenti sono progettati per soddisfare le diverse esigenze di studenti ed educatori, anche se il loro potenziale completo deve ancora emergere.

Diversi modelli di ITS supportano l'apprendimento personalizzato adattandosi alle esigenze individuali degli studenti. Essi identificano le lacune nell'apprendimento e forniscono feedback o contenuti su misura, rispondendo alle diverse abilità degli studenti, compresi quelli in difficoltà, ad alto potenziale o con bisogni educativi speciali. Inoltre, tali strumenti promuovono l'apprendimento autonomo e offrono agli educatori informazioni dettagliate sui progressi degli studenti, facilitando interventi mirati. Allo stesso modo, i sistemi LPG basati sull'IA assistono gli educatori attraverso la generazione di contenuti didattici differenziati e allineati al programma di studi, tenuto conto dei diversi livelli di competenza in materia di IA e delle preferenze degli studenti. Queste applicazioni si rivelano promettenti per il loro potenziale nel far risparmiare tempo agli educatori, nel supportare un insegnamento inclusivo e nel fornire strategie personalizzate in base al contesto educativo specifico.

Gli scenari descritti nelle sezioni successive illustrano il funzionamento di questi strumenti in contesti reali, mostrando come la XAI possa contribuire ad aumentare la trasparenza, promuovere la fiducia e migliorare il processo decisionale. In definitiva, la XAI offre un supporto significativo sia all'insegnamento che all'apprendimento. Prima di analizzare questi scenari, la sezione seguente propone una breve introduzione al formato delle spiegazioni, con l'obiettivo di evidenziare l'importanza di questo aspetto nell'ottica di facilitare la comprensione.

3.2 Spiegazioni visive

Attraverso l'adozione di misure volte a garantire la spiegabilità, i sistemi di IA possono diventare strumenti trasparenti e affidabili a supporto dello sviluppo educativo. Tuttavia, la complessità di questo processo risiede nel trovare un equilibrio tra le diverse prospettive degli stakeholder, ognuno dei quali nutre preoccupazioni ed aspettative specifiche sul modo in cui l'IA possa supportare efficacemente l'educazione.

Il formato in cui viene fornita la spiegazione agli utenti finali è un elemento fondamentale. Le principali modalità includono testo semplice, visualizzazioni ed esposizioni verbali (Minh et al, 2022; Johnson et al, 2023). Le evidenze scientifiche sottolineano l'efficacia delle spiegazioni visive (Sedrakyan et al, 2019; Bovek & Tversky, 2016). Queste rappresentazioni utilizzano interfacce grafiche per trasmettere informazioni attraverso immagini, grafici, diagrammi, plot, chart, reti e altre forme più astratte (Munzner, 2014; Sahin & Ifenthaler, 2021). Sebbene, ovviamente, le visualizzazioni possano includere anche testo, l'obiettivo in ambito educativo è evitare spiegazioni lunghe e descrittive, privilegiando invece contenuti più semplici e concettuali.

Negli ultimi anni sono stati compiuti notevoli progressi nel campo dei cruscotti grafici per l'istruzione (Sahin & Ifenthaler, 2021; Bull, 2020), e da essi la XAI può trarre beneficio. Per un'analisi approfondita delle visualizzazioni applicate alla XAI, si rimanda ad Alicioglu & Sun (2019) e, nel caso specifico dell'istruzione, a Ooge (2023). La loro rilevanza nel contesto di questo rapporto sarà illustrata attraverso due esempi reali, uno incentrato sugli studenti e l'altro sugli educatori. Entrambi riguardano sistemi di apprendimento potenziati dall'IA, ma l'obiettivo qui non è l'applicazione in sé, bensì mostrare le possibilità offerte dalle visualizzazioni per spiegazioni personalizzate e, di conseguenza, per una maggiore comprensibilità dello strumento.

Le immagini riportate nella figura 4 corrispondono a una piattaforma di e-learning basata sull'intelligenza artificiale per la scuola secondaria, che assegna esercizi adeguati al livello dello studente (Ooge, 2023):

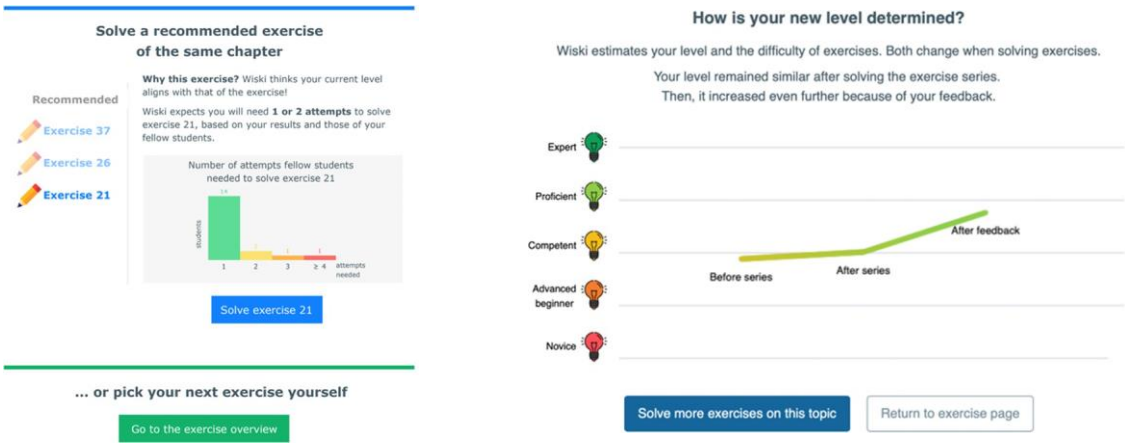

Figura 4: Piattaforma di e-learning basata sull'intelligenza artificiale che assegna esercizi agli studenti
Fonte: Ooge, 2023.

Nell'immagine a sinistra, il testo in alto spiega perché è stato selezionato un esercizio specifico (il n. 21), prima di fornire una giustificazione più dettagliata. Il grafico sottostante chiarisce la spiegazione utilizzando dati di gruppo e una rappresentazione a istogramma. L'immagine a destra presenta una visualizzazione dell'impatto delle scelte degli studenti dopo una serie di esercizi, con una spiegazione testuale sui progressi nella parte superiore e una spiegazione grafica in basso.

Per quanto riguarda gli educatori, le tre immagini seguenti, mostrate nella figura 5, corrispondono a Santa, un servizio di tutoraggio multipiattaforma per l'apprendimento dell'inglese basato sull'intelligenza artificiale:

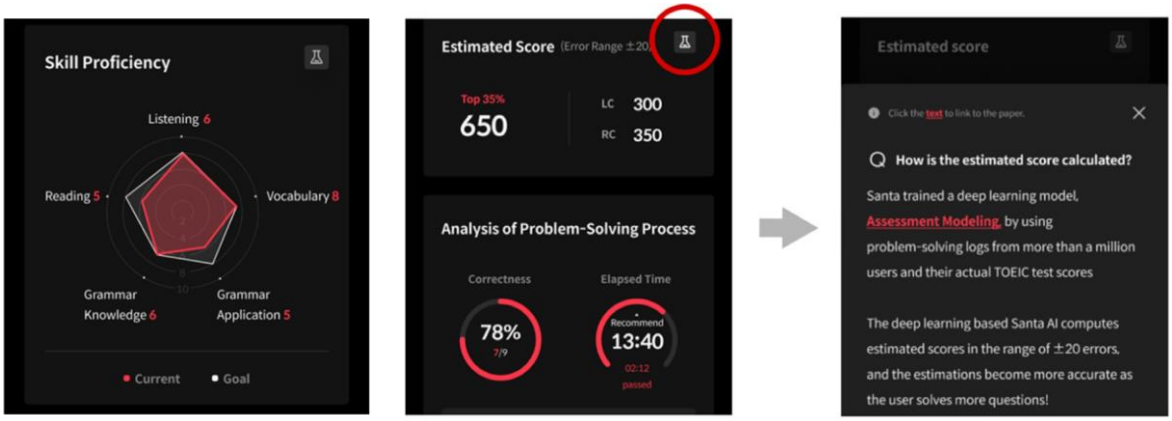

Figura 5: Spiegazioni dei punteggi stimati nel sistema di tutoraggio Santa.
Fonte: Kim et al, 2020.

L'immagine a sinistra mostra un grafico radar che illustra il livello di competenza dello studente nelle diverse prospettive di apprendimento analizzate dallo strumento. L'immagine centrale mostra un punteggio stimato ottenuto mediante una tecnica di deep learning (Kim et al, 2020). Quando l'insegnante clicca sull'icona in alto a destra, viene fornita una spiegazione del calcolo del punteggio, come si può vedere nell'immagine a destra. Con queste informazioni, l'insegnante dispone di maggiori dettagli sul livello di affidabilità del sistema di IA nella previsione del punteggio.

Questi due esempi evidenziano chiaramente il potenziale delle visualizzazioni grafiche nella XAI. Tuttavia, è anche evidente che resta ancora molto lavoro da fare per sviluppare spiegazioni che includano tutte le caratteristiche raccomandate sopra, nella tabella 3 e che siano utili sia per educatori che per studenti.

3.3 Caso d'uso 1: sistema di tutoraggio intelligente basato sull'IA

Per fornire agli utenti (*deployers* di IA) informazioni utili, i fornitori di IA devono includere caratteristiche di spiegabilità negli strumenti ITS. Tali integrazioni possono consentire agli utenti di convalidare le raccomandazioni fornite dal sistema di IA, intervenire in modo efficace e perfezionare il sistema per ottenere risultati migliori, garantendo che il sistema di IA realizzi il suo potenziale come strumento educativo trasformativo.

Scenario: applicazione dell'apprendimento adattivo potenziato dall'intelligenza artificiale a supporto dell'apprendimento della matematica nella scuola primaria

Una scuola primaria ha recentemente implementato un sistema di tutoraggio intelligente basato sull'IA (ITS) per supportare l'apprendimento della matematica da parte degli studenti. Il sistema si adatta alle capacità di ogni studente, fornendo materiali didattici personalizzati e feedback in tempo reale. Con l'adozione di tale strumento la scuola mira a migliorare i risultati di apprendimento, ottimizzare il carico di lavoro degli insegnanti e garantire che tutti gli studenti, indipendentemente dalle individuali esigenze di apprendimento, ricevano un sostegno adeguato. Il sistema offre agli insegnanti informazioni dettagliate sui progressi degli studenti attraverso una dashboard e fornisce report ai genitori e ai dirigenti scolastici. Ai sensi dell'AI Act, questo scenario sarebbe considerato un'applicazione ad alto rischio (articolo 6, paragrafo 2, in combinato disposto con l'allegato III, paragrafo 3, lettera b)), a causa del ruolo svolto dal sistema di IA nel processo decisionale, della personalizzazione basata sui dati e del potenziale impatto sui risultati di apprendimento degli studenti.

Emma, una studentessa di 10 anni, inizia a utilizzare il sistema per le sue lezioni di matematica. Inizialmente, Emma ha difficoltà con le frazioni e l'algebra di base, ma con l'ITS riceve compiti personalizzati e adattati al suo livello di competenza. Il sistema identifica le lacune di Emma nella comprensione della matematica e adatta la difficoltà degli esercizi in base alle prestazioni della studentessa. Visto dalla prospettiva dello studente, l'ITS basato sull'intelligenza artificiale è percepito come un tutor personale, che propone lezioni e attività adattate al suo livello di competenza. L'ITS fornisce suggerimenti in tempo reale e correzioni immediate per aiutare gli studenti a riflettere sugli errori e migliorare la propria conoscenza dei contenuti. Tuttavia, potrebbe non essere sempre chiaro perché vengono suggeriti determinati esercizi o in che modo avviene la valutazione delle competenze.

Prospettiva degli studenti

Quando comprendono il motivo per cui viene loro assegnato un determinato materiale, gli studenti sviluppano un senso di responsabilità e fiducia nelle raccomandazioni del sistema. Ciò è in linea anche con i principi relativi alla spiegabilità dell'IA e, in particolare, con l'obiettivo di promuovere la fiducia negli studenti. Per coinvolgere efficacemente gli studenti nel loro percorso di apprendimento e favorire un approccio all'apprendimento autodiretto è necessaria la trasparenza rispetto alla pertinenza e alla progressione degli esercizi assegnati**.**

Trasparenza e spiegabilità: per acquisire fiducia nell'ITS, gli studenti devono comprendere perché vengono assegnati determinati esercizi e come viene valutata la loro performance. La trasparenza e la spiegabilità su misura (adattando il tono e la complessità in base all'età) sono requisiti fondamentali affinché gli studenti comprendano la rilevanza e lo scopo di ogni compito e sviluppino un senso di controllo e fiducia nel sistema. È importante sapere come funziona l'algoritmo del sistema di raccomandazione e quali metodi vengono utilizzati per mantenere alta l'attenzione degli studenti (rischio di dark pattern).

Promuovere il coinvolgimento e il senso di responsabilità: quando gli studenti comprendono la logica alla base del loro percorso di apprendimento, è più probabile che si sentano coinvolti e motivati. Spiegare come i compiti vengano personalizzati e raccomandati in base alle loro esigenze specifiche aiuta gli studenti ad assumersi la responsabilità dei propri progressi e ad apprezzare il valore del sistema. (Maity & Deroy, 2024).

Equilibrio tra supporto e indipendenza: pur fungendo da guida di supporto, l'ITS deve incoraggiare gli studenti a sviluppare capacità autonome di risoluzione dei problemi. Riducendo gradualmente il livello di guida man mano che migliorano i risultati ottenuti, il sistema può aiutare a rafforzare la fiducia e l'autonomia degli studenti nell'affrontare sfide più complesse attraverso un sostegno adattivo (Liu et al, 2024). È necessario spiegare agli studenti e agli insegnanti come tale supporto possa essere bilanciato e se sia previsto un intervento umano nel processo (Ogata et al, 2024).

Integrità accademica: inoltre, è importante che gli studenti siano ben informati sull'uso appropriato ed etico degli ITS e sulle aspettative che dovrebbero avere, senza sminuire il ruolo dell'insegnante (Hong et al, 2022).

Apprendimento autonomo: un sistema ITS dovrebbe permettere agli studenti di stabilire obiettivi personali, riflettere sui propri progressi e personalizzare i percorsi di apprendimento in base ai propri interessi e bisogni. Integrare momenti di riflessione, collaborazione tra pari e consapevolezza etica coinvolge gli studenti in attività che promuovono competenze metacognitive e pensiero critico Majumdar et al, 2023).

Prospettiva dell'insegnante

Dal punto di vista dell'insegnante, è fondamentale che il sistema ITS basato sull'intelligenza artificiale possa essere valutato come efficace e affidabile rispetto agli approcci didattici e al design dell'apprendimento su cui è stato addestrato, così come per il supporto concreto che è in grado di offrire a tutti gli studenti, tenendo conto dei loro bisogni educativi speciali e dei diversi stili di apprendimento. Offrire un'educazione di qualità, che segua un approccio olistico e miri allo sviluppo intellettuale, emotivo e sociale dello studente, nel rispetto dei suoi diritti, è di primaria importanza.

Sorveglianza umana: gli sviluppatori e i fornitori di ITS devono garantire una supervisione umana integrata fino dalla fase di progettazione, in modo tale che gli insegnanti possano intervenire sulle decisioni dell'ITS e assegnare "manualmente" i compiti o apportare modifiche ai percorsi di apprendimento.

Apprendimento personalizzato e spiegazioni: i fornitori di ITS devono fornire spiegazioni riguardo all'approccio didattico e al design dell'apprendimento integrati nel sistema. È necessario chiarire come l'ITS operi per monitorare le prestazioni di ciascuno studente nei compiti, offrire suggerimenti per il miglioramento, stimoli per l'autovalutazione e commenti relativi allo stato emotivo.

Trasparenza e spiegabilità: la fiducia dell'insegnante nell'ITS può essere favorita da una comprensione approfondita delle funzioni dello strumento, così come del contesto educativo e formativo in cui questo è stato progettato. È inoltre fondamentale che lo strumento tenga conto delle diverse caratteristiche socio-culturali e di apprendimento della popolazione studentesca, come lingua, cultura e livello cognitivo legato all'età, per garantire che le conoscenze fornite siano pertinenti per tutti gli studenti (ethics by design). Tali informazioni dovrebbero essere accessibili all'insegnante dalla dashboard. Dal canto loro, gli insegnanti dovrebbero promuovere l'uso di ITS che includano le informazioni necessarie e che rispettino le dimensioni della XAI riportate nella tabella 4.

Integrità accademica:

L'ITS dovrebbe essere integrato nel processo educativo come un supporto per l'insegnante, non come un suo sostituto. È quindi importante che gli insegnanti informino e formino adeguatamente gli studenti riguardo sull'uso corretto ed etico del sistema, nonché sulle aspettative che gli studenti dovrebbero avere riguardo a quest'ultimo**.**

Rispetto dei diritti dei minori:

Infine, è fondamentale che l'insegnante si assicuri che l'ITS sia progettato nel rispetto dei diritti del minore, come la protezione dei dati personali e sensibili, la libertà di espressione e di scelta. L'uso di strumenti ITS non deve in alcun modo compromettere la sicurezza e il benessere degli studenti.

Prospettiva dei progettisti di programmi didattici

I progettisti di programmi didattici sono attori fondamentali nell'implementazione degli ITS per l'istruzione. Come già detto, questi sistemi basati sull'intelligenza artificiale hanno la capacità di migliorare l'apprendimento, personalizzando i contenuti e adattandosi alle esigenze del singolo. Tuttavia, la loro efficacia dipende dalla garanzia dell'allineamento con gli obiettivi del programma didattico, dall'offerta di pari opportunità nell'apprendimento e dalla risposta alle diverse esigenze degli studenti.

Comprendere le decisioni dell'ITS: sapere perché vengono assegnati compiti specifici e come questi siano in linea con gli standard curriculari. L'ITS dovrebbe avere una "modalità trasparenza" che mostri quali dati di input sono stati utilizzati (ad esempio, i punteggi dei test, i risultati ottenuti nei compiti precedenti) e percorsi logici che descrivano come i parametri di prestazione si collegano all'assegnazione dei compiti. Ad esempio, l'avvio di un'attività potrebbe fornire la seguente spiegazione: "Questo problema di geometria è stato selezionato perché lo studente ha dimostrato un'abilità dell'80% nelle competenze di algebra richieste".

Rilevamento dei bias: i fornitori di ITS dovrebbero consentire di identificare e mitigare eventuali diseguaglianze sistemiche nelle raccomandazioni o nell'assegnazione dei compiti. Gli utenti dovrebbero essere supportati da funzionalità di sistema che segnalino potenziali distorsioni, come la distribuzione non uniforme dei compiti di livello avanzato tra i generi o i gruppi socioeconomici. Ad esempio, un avviso di potenziale bias potrebbe indicare: "Le studentesse ricevono il 30% in meno di compiti avanzati rispetto ai loro colleghi maschi con prestazioni equivalenti".

Supportare la personalizzazione: i fornitori di ITS dovrebbero spiegare in che modo i contenuti vengono adattati agli studenti con difficoltà quali ADHD, dislessia o barriere linguistiche.

Garantire la trasparenza: i fornitori di ITS dovrebbero preparare indicazioni operative in grado di promuovere la fiducia tra insegnanti, genitori e dirigenti scolastici. Un modo efficace per raggiungere questo obiettivo potrebbe essere l'inclusione delle dimensioni della XAI riportate nella tabella 4, insieme alle funzionalità illustrate nella tabella 3.

Prospettiva dei leader educativi

La responsabilità principale di un leader educativo, come per esempio dirigente scolastico, consiste nel garantire l'implementazione mirata dell'ITS in linea con gli obiettivi educativi stabiliti, gli standard etici e le politiche vigenti. Ciò include la promozione di pari opportunità di apprendimento, il rafforzamento della fiducia delle parti interessate e la garanzia della conformità alle disposizioni nazionali e internazionali, come quelle menzionate nel capitolo precedente.

Allineamento agli obiettivi istituzionali e politici: l'implementazione di un ITS dovrebbe essere in linea con le priorità istituzionali, quali la cittadinanza digitale, i compiti a casa e le politiche relative al tempo trascorso davanti allo schermo. La spiegabilità garantisce che il sistema fornisca informazioni chiare su come vengono assegnati i compiti in conformità con tali obiettivi. Ad esempio, l'ITS dovrebbe mostrare in modo trasparente come adatta i compiti a casa per rispettare i limiti di tempo, sostenendo al contempo gli obiettivi di apprendimento.

Equità e accessibilità: i leader educativi dovrebbero garantire che l'ITS promuova pari opportunità di apprendimento e che si adatti alle diverse esigenze degli studenti. L'ITS dovrebbe spiegare la logica alla base delle raccomandazioni relative ai compiti, consentendo di identificare e affrontare eventuali distorsioni o disuguaglianze. Ad esempio, i rapporti generati dall'ITS dovrebbero spiegare chiaramente in che modo vengono personalizzati i compiti per gli studenti con disabilità o barriere linguistiche.

Spiegazioni personalizzate: promuovere gli ITS conformi alle dimensioni XAI illustrate nella tabella 4 consentirà agli studenti come Emma e ai loro insegnanti che utilizzano lo strumento di ricevere spiegazioni comprensibili.

Formazione e preparazione dei soggetti interessati: gli insegnanti e gli studenti dovrebbero ricevere una formazione adeguata, in modo tale da comprendere il funzionamento dell'ITS e il modo in cui questo influisce su di loro. In assenza di tale formazione, gli interessati potrebbero non avere fiducia nell'ITS o interpretarne erroneamente le funzionalità. Questo aspetto è trattato in modo più approfondito nel capitolo successivo.

Protezione dati e utilizzo etico: l'adesione a rigorose politiche di protezione dei dati richiede trasparenza nelle modalità di raccolta, archiviazione e utilizzo dei dati degli studenti da parte dell'ITS. Ad esempio, l'ITS dovrebbe indicare esplicitamente quali dati vengono raccolti, il loro scopo e in che modo essi supportano l'apprendimento personalizzato.

Monitoraggio e miglioramento continuo: i leader educativi dovrebbero valutare continuamente le prestazioni del sistema ITS. Lo strumento dovrebbe comunicare in che modo i nuovi algoritmi migliorano la regolazione della difficoltà dei compiti, permettendo di allineare tali aggiornamenti agli obiettivi istituzionali e alle aspettative degli stakeholder.

Prospettiva dei policy maker

Per i policy maker è essenziale che gli ITS nelle scuole primarie siano trasparenti, affidabili e utilizzati in modo responsabile. Questi strumenti possono influenzare in modo significativo le esperienze e lo sviluppo dei giovani studenti ed è pertanto è fondamentale disporre di linee guida chiare in materia di protezione dati, equità e responsabilità per garantirne un utilizzo etico.

Chiarezza nelle scelte di apprendimento adattivo: i policy maker dovrebbero richiedere ai fornitori di ITS di fornire spiegazioni chiare e accessibili riguardo agli adattamenti dei percorsi di apprendimento e all'assegnazione dei compiti. Una decisione trasparente garantisce che il percorso di apprendimento di ogni studente appaia intenzionale, riducendo la frustrazione e promuovendo un senso di responsabilità nei confronti dei propri progressi. A tal fine, i policy maker dovrebbero stabilire i requisiti per l'adozione di un approccio comune e completo in materia di spiegabilità, come suggerito nelle conclusioni del workshop della comunità EDEH sull'IA spiegabile nell'istruzione.

Tutela della riservatezza e sicurezza dei dati: data la delicatezza dei dati degli studenti più giovani, i sistemi ITS devono rispettare protocolli rigorosi per la raccolta, l'archiviazione e l'utilizzo dei dati. I policy maker dovrebbero richiedere ai fornitori di ITS di definire chiaramente quali dati vengono raccolti, il loro utilizzo previsto e i soggetti che vi hanno accesso.

Equità nelle raccomandazioni relative ai compiti: i sistemi di tutoraggio basati sull'intelligenza artificiale devono operare in modo equo, evitando distorsioni che possano favorire o svantaggiare determinati studenti. I policy maker dovrebbero introdurre verifiche periodiche sull'equità all'interno dei sistemi ITS, per garantire che gli adattamenti dell'apprendimento rimangano imparziali rispetto alle diverse origini e capacità degli studenti. Queste verifiche potrebbero includere l'analisi delle modalità di assegnazione dei compiti ai diversi gruppi di studenti, assicurandosi che gli adattamenti siano sempre personalizzati e giustificati.

Responsabilità e sorveglianza umana: strutture chiare di responsabilità sono fondamentali per gli ITS che regolano autonomamente i percorsi di apprendimento. I policy maker dovrebbero specificare chi è responsabile del monitoraggio di questi sistemi, chi risponde di eventuali danni e come viene garantita la supervisione umana e la possibilità di intervento. Ciò include, ad esempio, assicurarsi che insegnanti o dirigenti scolastici possano intervenire qualora le raccomandazioni generate dall'IA non rispondano adeguatamente alle esigenze degli studenti.

Promuovere l'alfabetizzazione all'IA per insegnanti e genitori: la trasparenza è maggiore quando genitori e insegnanti comprendono il ruolo dell'IA nell'apprendimento. I policy maker potrebbero sostenere programmi di formazione che consentano a genitori e insegnanti di interagire in modo consapevole con gli ITS, mettendoli in grado di mettere in discussione o adeguare le raccomandazioni, ove necessario.

Prospettiva degli sviluppatori

La generazione di un ITS per l'apprendimento della matematica nella scuola primaria richiede che lo sviluppatore presti particolare attenzione alla spiegabilità, all'adattabilità e all'equità per l'intero ciclo di vita dell'IA. L'obiettivo è quello di fornire uno strumento di apprendimento personalizzato, efficace e trasparente che soddisfi gli standard etici e supporti i risultati didattici.

Garantire la provenienza e l'integrità dei dati: gli sviluppatori devono stabilire solide pipeline di dati per gestire l'accuratezza e la pertinenza contestuale degli input, come le metriche delle prestazioni degli studenti e le cronologie di apprendimento. Utilizzando il tracciamento della provenienza dei dati e la convalida automatizzata, gli sviluppatori garantiscono l'integrità dei dati, mantenendo la conformità con framework come il GDPR dell'Unione Europea e il Family Educational Rights and Privacy Act (FERPA, Stati Uniti). I cruscotti di provenienza dovrebbero fornire informazioni in tempo reale sui flussi di dati, aiutando gli educatori a comprendere in che modo gli input influenzano le raccomandazioni sulle attività e rafforzando la fiducia nel sistema.

Elaborazione di raccomandazioni spiegabili: gli sviluppatori devono utilizzare tecniche XAI che consentano agli insegnanti di comprendere perché vengono assegnati compiti specifici. Essi devono inoltre rispettare le norme stabilite nel capitolo 2 in termini di trasparenza, cercando di applicare tecniche di IA che favoriscano l'interpretabilità.

Facilitare il feedback in tempo reale e l'adattabilità: l'adattabilità in tempo reale richiede sistemi decisionali robusti in grado di elaborare tempestivamente i dati provenienti dalle interazioni degli studenti. Le architetture basate sugli eventi e i modelli di raccomandazione sequenziale consentono all'ITS di regolare dinamicamente la difficoltà dei compiti. Gli sviluppatori devono creare cruscotti interattivi che mostrino alberi decisionali o diagrammi di flusso di adattamento per aiutare gli stakeholder dell'educazione a comprendere come vengono apportate le modifiche e consentire l'intervento manuale, ove necessario.

Equità e rilevamento delle distorsioni: gli sviluppatori devono integrare audit sull'equità e meccanismi di rilevamento delle distorsioni per garantire pari opportunità di apprendimento. Tecniche come i controlli di parità demografica e il *debiasing* avversariale possono essere utilizzate a tale scopo (Elazar & Goldberg, 2018) , per identificare e mitigare le distorsioni nelle raccomandazioni. Strumenti che visualizzano le tendenze demografiche, come le heatmap [mappe di calore o tracciamenti] delle assegnazioni dei compiti, aiutano gli sviluppatori a rilevare disparità e a perfezionare il sistema, in conformità alle linee guida etiche vigenti come quelle contenute nella Raccomandazione dell'UNESCO sull'etica dell'IA per i principi dell'educazione.

Integrazione del feedback degli utenti per un miglioramento continuo: i cicli di feedback sono essenziali per perfezionare l'ITS. Gli sviluppatori dovrebbero implementare interfacce per raccogliere i contributi degli stakeholder dell'educazione relativi alla pertinenza e all'efficacia delle attività, utilizzando algoritmi per l'elaborazione del linguaggio naturale (NLP) e per il clustering dei dati , al fine di analizzare le tendenze dei feedback. Gli sviluppatori dovrebbero utilizzare le informazioni emerse da questo processo per guidare il riaddestramento del modello e gli aggiornamenti del sistema, garantendo che l'ITS si evolva per rispondere ai bisogni del mondo reale.

Dimensioni chiave della XAI nel caso d'uso ITS

La tabella seguente illustra, per questo caso d'uso specifico, le dimensioni della XAI incluse nella tabella 4, di cui gli sviluppatori devono tenere conto nel progettare gli ITS. Le spiegazioni devono essere elaborate in base alle caratteristiche illustrate nella tabella 3 e adattate ai diversi soggetti interessati, al fine di garantire una corretta comprensione. Anche con esempi così generali è importante sottolineare la rilevanza di tutte le dimensioni.

Dimensione	Esempio
Ambito	**Livello globale**: illustrare le tendenze generali, ad esempio spiegando perché l'ITS rafforza le competenze in alcuni argomenti di algebra per la maggior parte degli studenti, sulla base dell'analisi del curriculum. **Livello locale**: motivare l'assegnazione di un esercizio specifico sulle frazioni a Emma, in relazione ai suoi errori precedenti e ai pattern ricorrenti nelle sue prestazioni.
Approfondimento	**Completo:** un rapporto dettagliato per gli insegnanti che mostra come i progressi di Emma in algebra siano migliorati nel tempo e quali fattori specifici abbiano contribuito. Selettivo: una breve spiegazione per i genitori di Emma su come l'ITS ha identificato la sua difficoltà con le frazioni e ha adattato i suoi compiti di conseguenza.
Alternative	**Spiegazioni contrastive:** spiegare perché Emma ha ricevuto esercizi sulle frazioni invece che sull'aritmetica di base, mostrando il suo divario di rendimento nelle frazioni. **Spiegazioni non contrastive:** mostrano i fattori utilizzati dall'ITS, come i punteggi bassi nei quiz, senza confrontare le alternative.
Flusso	**Condizionale:** "Se uno studente ottiene un punteggio inferiore al 70% negli esercizi sulle frazioni, consigliare ulteriori attività incentrate sulla comprensione concettuale". **Correlazionale:** mostrare come l'aumento del tempo dedicato alla pratica sia correlato al miglioramento dei punteggi di Emma nelle frazioni, aiutando gli educatori a comprendere i suoi progressi.

Tabella 10: Dimensioni della XAI nel caso d'uso ITS (esempi).

3.4 Caso d'uso 2: generatore di piani di lezione basato sull'intelligenza artificiale

Scenario: generazione di un piano di lezioni per la scuola media inferiore sulle frazioni

In una scuola media, gli insegnanti hanno recentemente iniziato a utilizzare un generatore di piani didattici basato sull'intelligenza artificiale (LPG) per supportare gli insegnanti nella preparazione delle lezioni. Lo strumento di IA analizza gli obiettivi del programma di studi, i risultati di apprendimento attesi e alcuni dati anonimizzati relativi alla classe in generale (come la descrizione generale del background e delle capacità degli studenti, senza dati specifici o personali).

Tuttavia, lo strumento è anche capace di utilizzare e analizzare dati degli studenti quali i risultati delle valutazioni recenti- e le loro preferenze, per esempio le attività pratiche o gli ausili visivi, il che aumenta il livello di rischio del sistema di IA, rendendo necessaria una valutazione prima che esso venga messo in uso.

Sulla base di tali input, il sistema genera un piano didattico personalizzato, utilizzando un modello di IA generativa che include una varietà di attività idonee a rispondere alle esigenze di apprendimento degli studenti. Il piano LPG suggerisce anche risorse digitali, il tempo stimato per ciascuna attività e opzioni di valutazione formativa. Gli insegnanti possono adattare il piano prima dell'implementazione per garantire che sia in linea con i loro approcci didattici e le esigenze della classe. La scuola mira a ridurre il carico di lavoro degli insegnanti, adattarsi alle diverse esigenze di apprendimento degli studenti e aumentare il coinvolgimento di tutti gli studenti in classe. Secondo l'AI Act, questo scenario (senza dati degli studenti) sarebbe considerato a basso o nullo rischio, poiché è l'insegnante a decidere sull'uso del piano di lezione generato dal LPG. Tuttavia, anche in casi a basso rischio come questo gli sviluppatori sono obbligati a garantire che il sistema rispetti i principi della XAI. Ciò include la progettazione del sistema in un modo che fornisca spiegazioni chiare su come vengono elaborati gli input e generati gli output. Tali spiegazioni sono necessarie per verificare se il testo generato è accurato e adeguato all'uso previsto. Inoltre, gli sviluppatori devono affrontare eventuali distorsioni nella generazione dei contenuti, per evitare la diffusione di stereotipi o supposizioni ingiuste. Garantendo trasparenza e permettendo verifiche di accuratezza tramite le spiegazioni fornite, gli sviluppatori contribuiscono a costruire fiducia e a dare agli insegnanti gli strumenti per utilizzare efficacemente e consapevolmente lo strumento.

La signora Lee, insegnante di scuola media, utilizza lo strumento LPG per creare un programma didattico sulle frazioni per la sua classe di matematica, in cui sono presenti studenti di livelli diversi. L'insegnante inserisce i dati anonimi relativi alle prestazioni della classe ricavati dalle valutazioni recenti e le sue preferenze in materia di apprendimento interattivo. Il sistema genera rapidamente un programma dettagliato che suggerisce attività in grado di coinvolgere gli studenti di diversi livelli. Lo strumento suggerisce giochi interattivi sulle frazioni per gli studenti che hanno difficoltà con i concetti matematici di base, compiti avanzati di problem solving per gli studenti più brillanti e lavori di gruppo per l'apprendimento collaborativo. Prima di finalizzare il programma, la signora Lee adatta una parte del gioco alla sua strategia didattica e modifica la portata del lavoro di gruppo in base al suo carico di lavoro. Per l'insegnante, lo strumento LPG è come un assistente personale, che crea lezioni coerenti con l'approccio pedagogico adottato e capace di coinvolgere gli studenti di diversi livelli. Tale strumento fornisce all'insegnante un supporto utile nella preparazione delle lezioni, sebbene la mancanza di spiegazioni chiare sul suo funzionamento potrebbe causare problemi di vario genere. Ad esempio, l'ambiguità nel processo decisionale potrebbe impedire alla signora Lee di capire perché vengono raccomandate determinate attività, limitando potenzialmente la sua capacità di adattare la lezione alle esigenze specifiche dei suoi studenti.

Prospettiva degli studenti: quando capiscono come funziona il sistema di intelligenza artificiale e quali risultati possono aspettarsi, gli studenti sviluppano fiducia nelle raccomandazioni dell'insegnante, che sono coerenti con i principi della XAI. Per questo motivo, è fondamentale che siano informati sull'uso e sulle funzionalità del sistema, sapendo che il loro coinvolgimento è assicurato e che i loro diritti sono tutelati.

Protezione dei dati: gli studenti e i genitori devono essere informati sull'eventuale uso dei dati degli studenti da parte di tali strumenti, sulle modalità di tale utilizzo e su altri aspetti della protezione dei dati garantiti dalle normative UE vigenti, tra le quali il GDPR. Il LPG utilizza solo dati anonimi relativi ai progressi dello studente, ma anche in questo caso è necessario fornire le informazioni opportune.

Differenziazione per studenti diversi: lo strumento può differenziare l'insegnamento suggerendo attività su misura per rispondere alle diverse esigenze degli studenti. Ad esempio, un piano di lezione può contenere suggerimenti per attività più motivanti per gli studenti in difficoltà, compiti avanzati per coinvolgere gli studenti più brillanti o lavori di gruppo per favorire la collaborazione tra studenti di diversi livelli. Tale approccio mirato può favorire opportunità di apprendimento eque, ma potrebbe anche generare effetti indesiderati, come l'ampliamento dei divari invece che l'aiuto reciproco. Il punto di forza della XAI in questo caso è che supporta il riesame e la rottura delle abitudini, nel caso in cui lo studente rilevi un ostacolo o un bias (Jauhiainen & Guerra, 2023), stimolando la riflessione collettiva e individuale sulle pratiche didattiche e sulle strategie di apprendimento.

Ruolo dello studente: anche se l'IA assiste l'insegnante nella creazione del piano e suggerisce attività e risorse coinvolgenti, è comunque necessario includere la supervisione umana, creando un'opzione che consenta agli studenti di commentare tali attività e risorse e fornire un feedback sull'adeguatezza delle attività alle loro esigenze.

Prospettiva dell'insegnante

Dal punto di vista dell'insegnante, è fondamentale che lo strumento LPG possa essere valutato come efficace e affidabile, sia per quanto riguarda le sue funzionalità, sia in relazione al contesto educativo e di apprendimento per cui è stato progettato. In particolare, considerando le diverse abilità degli studenti, è importante garantire che lo strumento di intelligenza artificiale valuti le prestazioni e le preferenze di ciascuno in modo imparziale, rispettando i diritti e i bisogni dello studente attraverso un approccio olistico.

Personalizzazione e adattabilità: lo strumento di IA può creare un piano didattico su misura, basato sugli obiettivi del programma scolastico, sulle preferenze didattiche dell'insegnante e, potenzialmente, sui dati relativi al rendimento degli studenti. Ciò aiuta a garantire che il piano risponda alle diverse esigenze di una classe eterogenea, allineandosi al contempo allo stile didattico dell'insegnante. Inoltre, è importante mantenere flessibilità, garantendo all'insegnante la possibilità di rivedere, adattare e/o modificare il programma in base alle dinamiche della classe. A tal fine, come discusso nel prossimo capitolo, è fondamentale sviluppare la capacità degli insegnanti di utilizzare l'IA generativa mantenendo autonomia e controllo.

Efficienza nella pianificazione: lo strumento di IA può semplificare il processo di pianificazione delle lezioni analizzando i dati e fornendo un piano strutturato, completo di risorse, tempistiche e opzioni di valutazione formativa. Ciò può far risparmiare tempo all'insegnante, consentendo di modificare il piano di lezione generato dall'IA e di renderlo più completo.

Autonomia e controllo dell'insegnante: sebbene lo strumento di intelligenza artificiale assista nella generazione del piano e suggerisca attività e risorse coinvolgenti, l'insegnante deve mantenere il controllo sulle decisioni finali. Questo equilibrio garantisce che lo strumento supporti il suo insegnamento senza sminuire la sua competenza professionale**.**

Apprendimento personalizzato e spiegazioni: lo strumento di IA può fornire spiegazioni relative all'approccio didattico e alla progettazione dell'apprendimento in base alle prestazioni di ciascuno studente nelle attività, suggerimenti per migliorare, spunti per l'auto monitoraggio e commenti relativi al livello emotivo. Inoltre, dovrebbe in qualche modo garantire la partecipazione equa e attiva di tutti gli studenti all'interno del gruppo, come parte dell'apprendimento collaborativo.

Integrità accademica: lo strumento di IA è stato integrato nel processo educativo per migliorare la dinamica della lezione, rendendola più attraente e fornendo un supporto aggiuntivo sia all'insegnante che agli studenti. È quindi importante che gli studenti siano ben informati e formati sull'uso appropriato ed etico dello strumento, così come sulle aspettative che possono avere rispetto a tale uso.

Trasparenza e spiegabilità: per garantire la fiducia dell'insegnante nello strumento di intelligenza artificiale, nel caso in esame, è importante che la professoressa Lee acquisisca una comprensione approfondita delle funzionalità di tale strumento, così come del contesto educativo e di apprendimento per cui è stato progettato. In particolare, considerando le diverse abilità degli studenti, è fondamentale che la professoressa Lee si assicuri che il sistema di IA valuti in modo imparziale e appropriato le prestazioni e le preferenze di ciascuno, al fine di incoraggiarli e motivarli a partecipare attivamente **(ethics by design).**

Rispetto dei diritti dei minori: è fondamentale che l'insegnante garantisca che il sistema di IA sia progettato nel rispetto dei diritti dei minori, quali la protezione dei dati personali e sensibili, la libertà di espressione e la libertà di scelta. L'utilizzo del sistema non può in alcun modo compromettere la sicurezza e il benessere degli studenti.

Prospettiva dei progettisti curricolari

Il ruolo dei progettisti curricolari i è quello di integrare efficacemente gli strumenti LPG basati sull'IA garantendo trasparenza, equità e inclusione. A tal fine, l'IA generativa deve essere utilizzata in modo appropriato, inquadrando i suoi risultati con link, testi o documenti da seguire, compreso il programma scolastico specifico. Concentrandosi sulla spiegabilità, questi strumenti possono migliorare i processi di pianificazione delle lezioni, preservando il controllo da parte degli insegnanti e dei progettisti.

Personalizzazione per classi con abilità miste: il sistema propone attività su misura per le diverse abilità, i livelli di coinvolgimento e le preferenze di apprendimento degli studenti. Senza una chiara motivazione delle raccomandazioni, i progettisti dei programmi di studio non sarebbero in grado di capire il motivo per cui compiti specifici vengano assegnati a determinati gruppi. Ad esempio, lo strumento LPG può assegnare compiti di addizione di numeri interi di base agli studenti in difficoltà, ma non spiegare i criteri utilizzati, generando sfiducia.

Allineamento agli standard: gli strumenti garantiscono che i piani di lezione rispettino i requisiti dei quadri educativi, dei curricoli o di altre linee guida stabilite dalle autorità scolastiche. Tuttavia, i progettisti devono attualmente convalidare manualmente gli output generati dall'IA, aumentando così il loro carico di lavoro e riducendo il loro ruolo creativo. Ad esempio, un progettista potrebbe modificare un'attività gamificata proposta per studenti con basso coinvolgimento, aggiungendo elementi collaborativi per favorire l'interazione tra pari.

Rilevanza culturale: il piano di lezione generato dall'IA adatta i contenuti per riflettere i contesti locali e culturali, migliorando il coinvolgimento e la capacità di identificazione degli studenti. I dati storici utilizzati dal sistema possono involontariamente perpetuare stereotipi, come l'assegnazione sproporzionata di compiti più semplici a specifici gruppi demografici.

Mitigazione delle distorsioni: algoritmi avanzati rilevano e segnalano potenziali distorsioni, promuovendo una equa distribuzione dei compiti. Gli strumenti di intelligenza artificiale dovrebbero individuare eventuali disuguaglianze nell'assegnazione dei compiti e suggerire alternative più eque. Ad esempio, gli avvisi di distorsioni possono segnalare ai progettisti se i compiti avanzati vengono assegnati in modo sproporzionato agli studenti maschi.

Efficienza: il sistema automatizza le attività ripetitive, consentendo ai progettisti di concentrarsi sul perfezionamento e sulla personalizzazione dei contenuti. Le raccomandazioni generiche mancano della flessibilità necessaria per adattarsi alle esigenze specifiche di ogni classe. Tuttavia, gli strumenti di intelligenza artificiale devono fornire esempi culturalmente rilevanti e garantire che i compiti siano adattati ai punti di forza individuali, senza rafforzare stereotipi. Ad esempio, un piano di lezione generato dall'IA può sostituire problemi generici con numeri interi con scenari che utilizzano variazioni locali di temperatura o dati di mercato, più vicini alla realtà di una classe eterogenea.

Prospettiva dei leader educativi

La responsabilità principale dei leader educativi, come per esempio i dirigenti scolastici, in relazione ai piani didattici basati sull'utilizzo dell'intelligenza artificiale, consiste nel garantire la trasparenza e la spiegabilità di tali strumenti. Questi aspetti sono fondamentali per promuovere la fiducia, consentire la personalizzazione e garantire l'allineamento con gli obiettivi didattici. I dirigenti devono affrontare le sfide e le esigenze specifiche di insegnanti, studenti e personale amministrativo per massimizzare il potenziale dello strumento.

Motivazioni alla base della personalizzazione delle lezioni: i leader educativi devono garantire che lo strumento di IA fornisca spiegazioni chiare su come vengono generati i piani didattici. Ad esempio, lo strumento dovrebbe spiegare perché consiglia giochi interattivi con le frazioni agli studenti in difficoltà e compiti di problem solving avanzati agli studenti più brillanti. Un processo di ragionamento trasparente garantisce che insegnanti e stakeholder comprendano e si fidino del processo di differenziazione, evitando scetticismi riguardo all'equità o all'efficacia dello strumento.

Allineamento con gli obiettivi e gli standard pedagogici: è fondamentale che lo strumento di intelligenza artificiale dimostri in che modo le attività suggerite siano allineate agli standard curricolari e alle priorità istituzionali. Ad esempio, i dirigenti scolastici devono poter verificare che lo strumento rispetti le aspettative specifiche per livello scolastico e gli obiettivi di apprendimento relativi alle frazioni. Senza trasparenza, c'è il rischio che i piani si discostino dai requisiti istituzionali, creando incoerenze nell'insegnamento in classe.

Supporto per classi con studenti di diversi livelli: i piani generati dall'IA devono mostrare in modo esplicito come soddisfano le diverse esigenze degli studenti, ad esempio fornendo attività pratiche per gli studenti cinestetici o un supporto per gli studenti in difficoltà. I dirigenti devono garantire la trasparenza dei criteri utilizzati per personalizzare i contenuti, al fine di garantire un accesso equo alle opportunità di apprendimento. Questa chiarezza consente agli insegnanti di intervenire in modo più mirato quando i piani non soddisfano le esigenze specifiche della classe.

Adattabilità alle preferenze degli insegnanti: i dirigenti devono garantire che gli insegnanti possano facilmente identificare e adattare le componenti suggerite dall'IA in modo che corrispondano al loro stile di insegnamento. Questa trasparenza garantisce che lo strumento migliori, anziché limitare, la flessibilità didattica, favorendo un maggiore coinvolgimento degli insegnanti nel sistema.

Efficacia delle risorse e delle valutazioni suggerite: i leader educativi dovrebbero valutare quanto efficacemente lo strumento di intelligenza artificiale giustifichi le sue raccomandazioni in ordine a risorse digitali, stime temporali e valutazioni formative. La trasparenza in questi ambiti garantisce che gli output siano concretamente utilizzabili e contestualmente rilevanti per gli insegnanti. Ad esempio, un piano che include un gioco digitale dovrebbe specificarne l'impatto atteso sugli esiti di apprendimento, permettendo ai dirigenti di valutare se tali strumenti siano coerenti con gli obiettivi istituzionali e pedagogici.

Prospettiva dei policy maker

Dal punto di vista dei policy maker, è essenziale che i piani di lezione generati dall'IA siano trasparenti, equi e utilizzati in modo responsabile nell'istruzione. Questi strumenti modellano il modo in cui le lezioni sono strutturate e impartite; è pertanto necessaria l'adozione di linee guida chiare in materia di protezione dati, equità, responsabilità e collaborazione per garantirne un utilizzo etico.

Trasparenza nelle raccomandazioni didattiche: i policy maker dovrebbero garantire che i piani di lezione generati dall'IA forniscano spiegazioni chiare e comprensibili delle loro raccomandazioni. Gli insegnanti devono sapere perché vengono suggerite determinate attività piuttosto che altre, soprattutto quando queste si diversificano per andare incontro alle esigenze degli studenti. Questa trasparenza permette agli insegnanti di fidarsi delle proposte dell'intelligenza artificiale e di interagire con esse in modo consapevole e sicuro. Ad esempio, l'IA potrebbe fornire una breve spiegazione del motivo per cui viene raccomandata un'attività di gruppo, evidenziando come essa supporti lo sviluppo delle competenze collaborative e aiutando così gli insegnanti a comprendere il ragionamento alla base della proposta.

Privacy e sicurezza dei dati degli studenti e degli insegnanti: poiché i piani di lezione generati dall'IA possono utilizzare sia i dati relativi al rendimento degli studenti che le preferenze degli insegnanti, si rende necessaria l'adozione di protocolli per il trattamento dei dati. Le politiche dovrebbero specificare quali dati vengono raccolti, come vengono conservati e chi può avervi accesso, in linea con il GDPR e con le altre disposizioni i vigenti in materia. I policy maker devono garantire che l'IA rispetti le pratiche di protezione dei dati, offrendo a insegnanti, studenti e genitori la sicurezza di un utilizzo sicuro ed etico delle informazioni, anche in sistemi a basso rischio.

Garantire l'equità nella personalizzazione delle lezioni: i piani di lezione generati dall'IA devono operare in modo imparziale, evitando distorsioni che potrebbero favorire determinati metodi di insegnamento o gruppi di studenti. I policy maker dovrebbero promuovere valutazioni periodiche dell'equità per confermare che l'IA adatti le lezioni in modo equo a studenti con background e capacità diversi. Ciò garantisce che tutti gli studenti beneficino di piani didattici adeguatamente personalizzati, creando un ambiente scolastico inclusivo.

Responsabilità e supervisione degli insegnanti: dato che i piani di lezione generati dall'IA guidano le attività in classe, è fondamentale disporre di misure chiare in materia di responsabilità. Le politiche dovrebbero definire chi è responsabile del monitoraggio del sistema e della revisione delle raccomandazioni, ove necessario. Gli insegnanti dovrebbero essere in grado di rivedere e adattare i piani per garantire che soddisfino gli obiettivi della classe. Ad esempio, se l'IA suggerisce compiti avanzati per un gruppo con abilità miste, gli insegnanti dovrebbero avere la possibilità di adeguare il piano in modo da andare incontro alle esigenze di tutti gli studenti.

Supportare l'alfabetizzazione all'IA degli insegnanti nella pianificazione delle lezioni: i policy maker possono promuovere la trasparenza sostenendo programmi di alfabetizzazione all'intelligenza artificiale che aiutino gli insegnanti a comprendere e utilizzare efficacemente i piani di lezione generati dall'IA. Una formazione che spieghi come l'IA formula le sue raccomandazioni permette agli insegnanti di interagire in modo critico con lo strumento, adattando i piani di lezione alle esigenze specifiche della propria classe. In tal modo si favorisce un approccio collaborativo, in cui l'IA agisce come una risorsa di supporto e non già come una sorta di comando.

Prospettiva degli sviluppatori

Dal punto di vista dello sviluppatore, nella creazione di un piano di lezione generato dall'IA per insegnanti della scuola secondaria di primo grado, come la professoressa Lee, l'obiettivo è sviluppare uno strumento che supporti una pianificazione didattica adattabile, spiegabile ed equa lungo tutto il ciclo di vita dell'intelligenza artificiale. Lo scopo è garantire che lo strumento sia trasparente, flessibile e imparziale, offrendo agli insegnanti la possibilità di proporre lezioni coinvolgenti e connesse al mondo reale, in linea con gli standard normativi ed etici.

Raccomandazioni spiegabili per la pianificazione delle lezioni: gli insegnanti devono potersi fidare delle raccomandazioni generate dall'IA. Strumenti come SHAP o LIME possono chiarire perché vengono suggerite attività specifiche, come collegare un'attività relativa alla siccità alla richiesta dell'insegnante di applicazioni ambientali in matematica. Questa trasparenza è in linea con il "diritto alla spiegazione" previsto dall'Unione Europea e con gli standard etici dell'IEEE e garantisce che gli insegnanti comprendano e si fidino del ragionamento dell'intelligenza artificiale.

Raccolta ed elaborazione dei dati contestualmente rilevanti: gli sviluppatori devono garantire che lo strumento integri gli standard curriculari, i risultati precedenti degli studenti ed esempi reali (ad esempio, scenari di cambiamento climatico) per elaborare piani didattici efficaci. Ad esempio, l'utilizzo dei dati sulla siccità in una lezione di matematica sulle frazioni aiuta a contestualizzare concetti astratti. Tuttavia, bilanciare l'integrità e la pertinenza dei dati è una sfida. Il rispetto del GDPR e degli standard etici garantisce la trasparenza nell'approvvigionamento dei dati, mentre i sistemi di tracciamento della provenienza visualizzano l'impatto dei dati contestuali sui suggerimenti didattici.

Feedback in tempo reale e adeguamenti basati sul contesto: il generatore deve consentire adeguamenti immediati sulla base di metriche di coinvolgimento. Ad esempio, se gli esercizi statici perdono l'interesse degli studenti, l'IA potrebbe suggerire di passare a una simulazione interattiva di un'alluvione. La trasparenza nella progettazione dell'applicazione in merito al modo in cui i dati sul coinvolgimento influenzano questi adeguamenti è in linea con i principi dell'OCSE sull'IA e consente agli educatori di gestire le lezioni in modo dinamico ed efficace.

Rilevamento delle distorsioni e verifica dell'equità: garantire un apprendimento equo richiede l'identificazione e la correzione dei bias presenti nei dati o negli algoritmi. Verifiche periodiche sull'equità possono evidenziare disparità, come ad esempio un'eccessiva enfasi su esempi urbani o rurali. In linea con i principi etici dell'IA elaborati dall'UNESCO, gli sviluppatori possono implementare strumenti per monitorare e adeguare la distribuzione dei contenuti, garantendo l'inclusività tra studenti con background diversi.

Feedback degli utenti e miglioramento iterativo: il perfezionamento continuo basato sul feedback degli insegnanti è fondamentale. Un semplice meccanismo di feedback consente ai docenti di valutare le attività suggerite dall'IA, ad esempio la pertinenza di una lezione di matematica legata a un'alluvione. Le informazioni raccolte attraverso questo processo guidano gli sviluppatori nel migliorare l'adattabilità dello strumento e il suo allineamento con il contesto, incarnando un approccio "human-in-the-loop" (l'uomo al centro del processo) in conformità con gli Orientamenti etici della Commissione europea.

Dimensioni chiave della XAI nel caso d'uso relativo al piano di lezione generato dall'IA (LPG)

La tabella seguente illustra, per questo secondo caso d'uso, le diverse dimensioni della XAI incluse nella tabella 4. Anche in questo caso, le spiegazioni riportate nella tabella devono essere elaborate in base alle caratteristiche descritte nella tabella 3 e adattate ai diversi soggetti interessati per essere comprese correttamente. Gli sviluppatori devono tenere conto di queste dimensioni durante la progettazione del loro LPG.

Dimensione	Esempio
Ambito	**Globale:** spiegare perché il piano di lezione generato dall'IA dà priorità alle attività pratiche per determinati argomenti, sulla base di obiettivi curricolari generali. **Locale:** spiegare perché è stato raccomandato un supporto visivo specifico per la lezione sulle frazioni dell'insegnante, dati i risultati ottenuti dai suoi studenti.
Approfondimento	**Completo:** fornire all'insegnante una spiegazione dettagliata di come il LPG combina gli standard curricolari, i dati di valutazione e le preferenze per generare piani. **Selettivo:** una breve nota in cui si spiega perché lo strumento ha suggerito il lavoro di gruppo per gli studenti in difficoltà.
Alternative	**Spiegazione contrastiva:** evidenziare perché il LPG ha suggerito un gioco visivo sulle frazioni invece di una lezione frontale, sulla base della preferenza della signora Lee per l'apprendimento interattivo. **Spiegazione non contrastiva:** elenca i fattori principali (ad esempio, i punteggi di coinvolgimento degli studenti) presi in considerazione senza confrontare le alternative.
Flusso	**Condizionale:** "Se gli studenti ottengono nelle frazioni risultati inferiori al livello previsto, si suggerisce di includere un'attività di revisione prima di introdurre nuovi concetti".

Tabella 11: Dimensioni della XAI nel caso d'uso di un LPG (esempi)

3.5 Livello di intervento degli stakeholder e punti di attenzione

Dai due scenari descritti in precedenza emerge con chiarezza il ruolo centrale della spiegabilità nel garantire che strumenti come ITS o LPG siano efficaci, affidabili e utilizzabili in contesti educativi reali. Sebbene tali strumenti mirino a migliorare le esperienze di apprendimento, affrontare le sfide e personalizzare l'insegnamento, il loro pieno potenziale dipende dalla capacità di fornire spiegazioni chiare, in grado di garantire al contempo l'accuratezza e la coerenza con gli obiettivi educativi. La seguente tabella riassume i livelli di intervento e i punti chiave di attenzione per i principali stakeholder dell'istruzione, ottenuti dall'analisi dei due casi d'uso precedenti, con l'obiettivo di fornire indicazioni utili ai lettori che facciano uso di strumenti simili in ambito educativo.

Stakeholder	Esempio	Punti di attenzione
Studenti	• Coinvolgimento diretto degli studenti con contenuti personalizzati, attività mirate e assegnazione di compiti adeguati al loro livello. • Partecipazione attiva e feedback immediato per favorire l'apprendimento autonomo e l'iniziativa personale.	• Spiegazioni chiare e personalizzate con dashboard accessibili. • Trasparenza nell'assegnazione dei compiti e nel processo decisionale dell'IA. • Trattamento etico dei dati e supporto all'autonomia degli studenti.
Insegnanti	• Sorveglianza e convalida delle raccomandazioni, dei piani delle lezioni e delle attività generate dall'IA. • Coinvolgimento attivo nell'adattare i risultati dell'IA al contesto della classe e nell'individuare eventuali distorsioni.	• Garantire equità, inclusività e coerenza con gli obiettivi di apprendimento. • Mantenere un controllo etico con risultati modificabili e standard chiari. • Promuovere l'alfabetizzazione all'IA e correggere i risultati distorti attraverso l'intervento manuale.
Progettisti curricolarii	• Allineamento e controllo della qualità dei risultati dell'IA con gli standard curriculari stabiliti. • Monitoraggio e correzione dei contenuti generati dall'IA per individuare eventuali distorsioni e garantire la coerenza con le linee guida istituzionali.	• Porre l'accento sulla trasparenza nella creazione dei contenuti e sull'allineamento con il curriculum. • Correggere le distorsioni e promuovere equità, diversità e inclusione. • Supportare meccanismi di feedback interattivi e integrare l'alfabetizzazione all'IA nella progettazione curricolare.
Leader educativi	• Supervisionare l'implementazione dei sistemi di IA all'interno dell'istituto, garantendo la conformità alle politiche educative. • Facilitare lo sviluppo professionale e monitorare le prestazioni complessive dell'IA.	• Garantire la trasparenza e il rispetto dell'etica attraverso cruscotti istituzionali e tracce audit. • Mantenere l'equità e l'accessibilità. • Supportare la formazione continua degli insegnanti e i cicli di feedback con gli stakeholder.
Policy maker	• Garantire la supervisione da parte degli organismi competenti e l'equità nel processo decisionale dell'IA. • Garantire che i sistemi di IA funzionino in modo trasparente e che siano in linea con gli obiettivi più ampi delle politiche pubbliche.	• Salvaguardare la protezione dei dati e i diritti dei cittadini. • Imporre obblighi di responsabilità, audit di equità e gestione del rischio nell'implementazione dell'IA. • Promuovere l'alfabetizzazione dei cittadini all'IA e gli standard etici nei sistemi
Sviluppatori	• Sviluppare sistemi di IA che forniscano raccomandazioni personalizzate integrando tecniche XAI. • Creare solide pipeline di dati con tracciabilità della provenienza e dashboard in tempo reale, consentendo adattamenti dinamici. • Facilitare i cicli di feedback in tempo reale e le sostituzioni manuali, per perfezionare le raccomandazioni sulla base delle interazioni dirette con gli studenti e degli input dei docenti.	• Garantire l'accuratezza e l'integrità dei dati, nel rispetto del GDPR, del FERPA e di altri quadri normativi pertinenti. • Incorporare regolari audit di equità e meccanismi di rilevamento delle distorsioni supportati da strumenti visivi. • Mantenere output chiari e interpretabili insieme a spiegazioni contestualizzate che generino fiducia tra gli stakeholder del settore dell'istruzione. • Utilizza il feedback degli stakeholder, analizzato per guidare il miglioramento continuo del sistema.

Tabella 12: Punti chiave di attenzione e interazioni con il sistema di IA per i principali stakeholder dell'istruzione.

3.1. Garantire la spiegabilità incentrata sulla persona nell'applicazione dell'IA in ambito educativo: ruoli, responsabilità e necessità di supervisione

Garantire la spiegabilità dell'intelligenza artificiale in ambito educativo è compito che va oltre gli algoritmi, poiché richiede un impegno attivo, una partecipazione etica e una responsabilità condivisa. Gli educatori svolgono un ruolo fondamentale nell'interpretare e contestualizzare gli output dell'IA per allinearli agli obiettivi di apprendimento individuali, mentre gli studenti si assumono la responsabilità di un uso corretto degli strumenti di IA, dimostrando la propria integrità accademica attraverso la capacità di spiegare come interagiscono con questi ultimi. I genitori e le autorità scolastiche traggono beneficio da cruscotti esplicativi che garantiscono trasparenza nel rispetto della protezione dei dati. I policy maker si basano su audit di equità e valutazioni delle distorsioni per promuovere l'equità, mentre agli sviluppatori spetta il compito di integrare feedback continui per realizzare strumenti trasparenti, adattabili e coerenti con i bisogni degli stakeholder.

L'integrazione del principio di "spiegabilità by design" nei sistemi di intelligenza artificiale per l'istruzione non è compito esclusivamente degli sviluppatori, ma ricade anche sulle imprese, attraverso le scelte aziendali e istituzionali, coinvolgendo anche i fornitori e i distributori, i quali devono garantire la conformità dei sistemi alle regolamentazioni del mercato e alle esigenze degli utenti finali. Naturalmente, un ruolo chiave spetta anche ai policy maker, che hanno il compito di facilitare l'integrazione delle prospettive educative richieste nei processi di sviluppo.

Pur essendo fondamentale il miglioramento degli aspetti tecnici della spiegabilità, con particolare attenzione alla trasparenza e all'interpretabilità dei modelli, la supervisione umana resta imprescindibile per garantire che gli strumenti di intelligenza artificiale siano significativi, affidabili e concretamente utilizzabili. Gli algoritmi specifici di XAI non sono, da soli, in grado di rispondere pienamente alla complessità dei contesti educativi reali, poiché spesso mancano di adattabilità e profondità.

Il giudizio umano è indispensabile per validare, interpretare e contestualizzare gli output dell'IA, assicurandone la trasparenza, il rispetto dei principi etici e l'allineamento con i bisogni educativi.

Rendere effettiva la spiegabilità dell'intelligenza artificiale in ambito educativo richiede un ecosistema collaborativo, in cui educatori, studenti, genitori, sviluppatori e policy maker siano attivamente coinvolti in un dialogo continuo. Siffatto impegno deve andare oltre le soluzioni puramente tecniche, ponendo l'attenzione su un design centrato sulla persona e su una responsabilità condivisa. In tale scenario, ciascun attore ricopre un ruolo fondamentale, con l'obiettivo comune di creare un sistema adattivo e riflessivo, che valorizzi il giudizio umano, salvaguardi l'integrità accademica e garantisca che gli strumenti di IA restino mezzi di supporto, in grado di arricchire la comprensione educativa.

Attraverso la promozione di una comunicazione aperta, di meccanismi di feedback regolari e di un impegno verso l'innovazione etica, la comunità degli stakeholder può orientare lo sviluppo di tecnologie basate sull'IA verso strumenti che siano trasparenti, sensibili al contesto e coerenti con i diversi bisogni educativi.

Le competenze degli educatori in relazione alla XAI

4.1 Contesto

Il capitolo introduttivo ha evidenziato come le tecniche e le procedure sviluppate nell'ambito della XAI siano orientate all'utente finale, con l'obiettivo di costruire un adeguato livello di fiducia nei sistemi di intelligenza artificiale.

Gli utenti finali possono includere una varietà di stakeholder nel contesto dell'IA, dagli sviluppatori al pubblico generale, poiché tutti necessitano di un certo livello di comprensione degli output generati dal sistema. Tuttavia, nell'ambito dell'istruzione, il principale destinatario sono gli studenti.

Gli studenti devono essere formati per vivere in un mondo permeato dall'intelligenza artificiale e, più nello specifico, per utilizzare tali sistemi a fini di apprendimento, valorizzando la propria autonomia e sviluppando il pensiero critico.

Di conseguenza, il primo e più rilevante gruppo di stakeholder da considerare quando si parla di spiegabilità dell'IA nel contesto educativo è quello degli insegnanti, in quanto responsabili della formazione degli studenti in questa prospettiva.

Nella Raccomandazione dell'UNESCO sull'etica dell'intelligenza artificiale, si legge che "L'alfabetizzazione e la consapevolezza sull'intelligenza artificiale sono fondamentali per tutti i cittadini, affinché possano orientarsi e interagire in modo responsabile con i sistemi di IA" (UNESCO, 2022, p. 36- *Traduzione libera*). Ciò significa fornire a studenti e insegnanti non solo conoscenze tecnologiche o competenze tecniche, ma anche la capacità di mettere costantemente in discussione e analizzare criticamente i sistemi di IA, al fine di adottare tali strumenti potenti come opportunità, comprendendo al contempo le implicazioni etiche, i limiti e i potenziali rischi associati ai sistemi di intelligenza artificiale.

Come spiegato nel capitolo 2, l'AI Act sottolinea l'importanza della trasparenza e della responsabilità, garantendo che i sistemi di IA siano progettati e implementati con linee guida chiare per il loro utilizzo e i potenziali rischi (articolo 3, dell'AI Act). Quando si parla di educare gli stakeholder all'IA, è fondamentale includere la comprensione di come questi sistemi possano talvolta rafforzare i pregiudizi, creare dilemmi etici e avere un impatto sociale di vasta portata, come illustra il framework DigComp. Senza una comprensione più approfondita, gli studenti e gli educatori potrebbero non disporre degli strumenti necessari per orientarsi nelle complessità dell'IA, con il rischio di un uso improprio o di un'accettazione acritica della tecnologia.

Quest'ultimo capitolo è incentrato sulle competenze che gli educatori devono possedere in materia di IA per quanto riguarda la XAI, sia per utilizzare strumenti basati sull'IA nella loro attività didattica, sia per insegnare i fondamenti tecnici e le implicazioni etiche di questa tecnologia. Utilizzando come riferimento il "Quadro di Competenze per l'Intelligenza Artificiale per Studenti" dell'UNESCO (Fengchun & Kelly, 2024), la successiva sezione 4.3 evidenzia le competenze fondamentali degli educatori che sono direttamente correlate alla XAI, secondo i livelli di classificazione internazionale dell'istruzione ISCED. Si apre tuttavia un nuovo scenario: indipendentemente dal livello di istruzione, gli educatori saranno chiamati a selezionare lo strumento di IA più appropriato in funzione degli studenti, del contesto specifico e degli obiettivi di apprendimento. Pertanto, oltre alle competenze di base in materia di IA per la XAI, l'educatore dovrà possedere le competenze specifiche illustrate di seguito nella sezione 4.4, per poter valutare le caratteristiche di spiegabilità degli strumenti sulla base delle dimensioni chiave definite nella tabella 4.

Se mettono in pratica tutte queste competenze, gli educatori saranno in grado di analizzare e valutare qualsiasi strumento di XAI, di utilizzarlo e sfruttarne appieno le potenzialità, creando esperienze di apprendimento adattate al proprio contesto specifico. Essi avranno inoltre la capacità di trasmettere agli studenti le competenze fondamentali per sviluppare un pensiero critico nei confronti dei sistemi di intelligenza artificiale, aumentando così la loro autonomia e sicurezza.

Esempi rappresentativi di possibili applicazioni della XAI nei diversi livelli ISCED sono presentati nella sezione 4.5, a conclusione di questo capitolo, insieme ad alcune raccomandazioni, rivolte ai principali stakeholder nell'ambito della alfabetizzazione all'intelligenza artificiale.

4.2 Principi fondamentali dell'IA e loro connessione con la XAI

Il pensiero critico come obiettivo

In un'era caratterizzata dall'intelligenza artificiale, le considerazioni e i dilemmi di carattere etico richiedono più che mai un pensiero critico. Pertanto, gli studenti, insieme ai loro educatori, devono essere protetti dall'atrofia cognitiva o dalla manipolazione, come quelle poste dall'IA generativa, e devono essere dotati di atteggiamenti di pensiero critico, quali profondità intellettuale, ragionamento basato su dati, informazioni e prove, nonché fiducia nella ragione, come delineato nel Paul-Elder Critical Thinking Framework (Paul & Elder, 2006). La trasparenza e la spiegabilità dei sistemi di IA sono prerequisiti per esercitare il pensiero critico. Senza accesso a dati, modelli e algoritmi, la capacità umana di esercitare un controllo su di essi risulta compromessa.

I quadri di riferimento dell'UNESCO sulle competenze in materia di IA riconoscono il pensiero critico come una competenza fondamentale per educatori e studenti, soprattutto nel contesto dell'integrazione dell'IA nei sistemi educativi. I quadri promuovono un approccio solido e responsabile, volto a garantire che l'uso dell'IA nell'istruzione non sia solo efficace, ma anche trasparente e responsabile, rafforzando così la fiducia tra tutte le parti coinvolte. In questo contesto si inserisce anche il Transparency Index Framework (Chaudhry et al., 2022), strettamente connesso al pensiero critico. Questo strumento favorisce una maggiore comprensione e un processo decisionale più consapevole tra gli stakeholder, grazie alla trasparenza di dati e algoritmi. Tale chiarezza permette agli utenti di valutare in modo critico le implicazioni degli strumenti di IA nel contesto educativo, promuovendo una cultura dell'indagine e stimolando domande critiche sui sistemi utilizzati.

Un esempio illuminante che evidenzia l'urgenza di affrontare i rischi legati all'opacità di alcuni sistemi di intelligenza artificiale è rappresentato dal MIT AI Risk Repository, in cui si sottolinea come gli algoritmi con processi decisionali opachi possano generare bias involontari o risultati discriminatori. Tali conseguenze rischiano di compromettere le fondamenta stesse dei sistemi educativi e la loro missione principale di *educere*, ovvero di far emergere e sviluppare i talenti degli studenti attraverso varie forme di alfabetizzazione, tutte radicate nel pensiero critico come valore universale. Per questo motivo, è fondamentale intervenire tempestivamente nei sistemi educativi, promuovendo trasparenza e responsabilità nell'adozione dell'IA, consentendo sia agli studenti che agli educatori di adottare l'IA come una vera opportunità per sviluppare ulteriormente il pensiero critico. Questo, a sua volta, dovrebbe tradursi in approcci basati sulla ricerca e sulla risoluzione di problemi, piuttosto che in una passiva accettazione dei risultati generati dall'IA a partire da semplici suggerimenti che, sempre più spesso, mancano di chiarezza e di una tracciabilità verificabile.

AI LIteracy: la strada verso un uso consapevole dell'intelligenza artificiale

Una delle definizioni più citate di alfabetizzazione all'IA è "un insieme di competenze che consente agli individui di valutare criticamente le tecnologie di IA, comunicare e collaborare efficacemente con l'IA e utilizzare l'IA come strumento online, a casa e sul posto di lavoro" (Long & Magerko, 2020, p. 2). In altre parole, essere alfabetizzati all'IA significa essere in grado di utilizzare, monitorare e riflettere criticamente sugli strumenti di IA in contesti personali, professionali o educativi. Di conseguenza, è un modo per raggiungere l'obiettivo del pensiero critico stabilito sopra.

Nel caso specifico, ma molto rilevante, dell'IA generativa, l'alfabetizzazione viene considerata una competenza essenziale, accanto alle tradizionali competenze digitali, per aiutare gli studenti a personalizzare il proprio apprendimento (gruppo di lavoro AI, 2023AAIN Generative). In primo luogo, è necessario comprendere i diversi tipi di strumenti di IA generativa, che possono creare contenuti come testi e immagini, imparare a formulare prompt efficaci per ottenere i risultati desiderati e utilizzarli per migliorare l'apprendimento e il

lavoro. In secondo luogo, la capacità di valutare l'accuratezza e l'affidabilità dei risultati ottenuti include l'identificazione di potenziali bias e informazioni false, oltre a richiedere la verifica delle informazioni fornite dall'IA con risorse affidabili. In terzo luogo, comprende l'uso etico e responsabile dell'IA, il che significa proteggere i dati sensibili, riconoscere questioni come la protezione dei dati e riconoscere l'uso dell'IA nel lavoro accademico (Pretorius, 2023).

L'alfabetizzazione all'intelligenza artificiale è fondamentale per la XAI sotto diversi aspetti. Poiché si basa sulla comprensione delle decisioni prese dall'IA, la XAI richiede agli utenti la capacità di cogliere il "perché" e il "come" del comportamento di un sistema di IA. Inoltre, la capacità di valutare l'affidabilità delle informazioni fornite dall'IA può essere notevolmente rafforzata dall'introduzione di processi trasparenti, in grado di rivelare potenziali distorsioni o errori. La XAI, infatti, permette di interrogarsi sugli aspetti etici legati all'IA, rendendo visibili i bias, i vincoli e i limiti dei processi decisionali e offrendo in tal modo agli utenti gli strumenti per valutare criticamente la tecnologia. Infine, la XAI è perfettamente in linea con i principi dell'AI literacy, in quanto rende l'IA più accessibile anche a chi non possiede competenze tecniche avanzate, permettendo a un pubblico più ampio di interagire in modo responsabile con queste tecnologie.

4.3 Competenze e principi fondamentali per l'integrazione della XAI nell'istruzione

Con l'obiettivo di fornire un percorso chiaro per integrare la XAI nell'istruzione, la presente sezione si concentrerà sulle competenze fondamentali per gli insegnanti in questo ambito, che saranno sviluppate sulla base del *Quadro delle competenze in materia di IA dell'UNESCO per gli studenti* (Fengchun & Kelly, 2024). Questo quadro è stato scelto come riferimento perché include competenze sia per apprendere con l'IA sia per apprendere sull'IA (Commissione Europea, 2023), mentre il quadro per gli insegnanti. è focalizzato sulle competenze per insegnare con l'IA. Pertanto, dal quadro per gli studenti sono state selezionate le competenze relative alla XAI e sono state definite quelle richieste agli educatori in base a esse. Le competenze sono state raggruppate in base alla loro in base alla loro rilevanza per educatori di diversi livelli ISCED. Esse progrediscono attraverso i livelli, partendo dalle competenze fondamentali di base per i livelli ISCED inferiori, fino ad arrivare a competenze avanzate e più specifiche per i livelli superiori. Si presume che i *livelli superiori includano le competenze acquisite nei livelli precedenti, quindi verranno evitate le ripetizioni.*

Livelli ISCED 1-3 (istruzione primaria, istruzione secondaria di primo grado, istruzione secondaria di secondo grado)

Le seguenti competenze riuniscono principi e abilità che permettono agli studenti, a questi livelli iniziali, di comprendere, valutare e ottimizzare i sistemi di intelligenza artificiale, garantendo che operino in modo trasparente e responsabile — un aspetto essenziale della XAI.

Etica dell'IA: ai livelli di comprensione, applicazione e progettazione, viene sottolineata l'importanza di integrare i principi etici in ogni fase del ciclo di vita dell'IA, dalla concettualizzazione all'implementazione. Trasparenza e spiegabilità svolgono un ruolo fondamentale in questo processo, consentendo a un pubblico informato di partecipare attivamente alla regolamentazione e all'uso etico dell'IA. In tale contesto, la XAI è essenziale, poiché consente di analizzare e giustificare come e perché l'IA prenda determinate decisioni. Ciò garantisce che le decisioni siano allineate agli standard etici e aiutino a prevenire pratiche discriminatorie o distorsioni. Inoltre, la XAI favorisce la comprensione dei suoi processi da parte del pubblico, permettendo ai cittadini di adottare una prospettiva critica e di prendere decisioni consapevoli riguardo alla sua adozione e al suo uso responsabile.

Approccio antropocentrico a livello di applicazione: questa competenza implica la consapevolezza che le persone sono responsabili delle decisioni generate dall'IA, soprattutto nei contesti ad alto impatto. La XAI fornisce una base per giustificare tali decisioni e consente sia ai progettisti sia agli utenti di assumersi le responsabilità legali ed etiche connesse. Questa mentalità è strettamente collegata ai livelli ISCED 1, 2 e 3, poiché sviluppa una consapevolezza di base della responsabilità umana nell'IA, partendo dall'esplorazione delle applicazioni dell'IA nel mondo reale fino ad arrivare all'importanza della spiegabilità per garantire la responsabilità etica e legale.

Tecniche e applicazioni dell'IA a livello di comprensione e progettazione: queste competenze comprendono la capacità di determinare quando e come l'IA dovrebbe essere utilizzata in modo appropriato. La XAI consente di valutare se i modelli e le architetture selezionati sono quelli più adatti al problema da risolvere, offrendo una visione dettagliata dei limiti e delle capacità dei modelli. Grazie a questo approccio, gli studenti del livello ISCED 1 vengono introdotti alle applicazioni di base dell'IA per comprendere dove e come l'IA viene utilizzata nella vita quotidiana. Al livello ISCED 2, le competenze servono a esplorare quando l'IA dovrebbe essere applicata e analizzano la sua idoneità per compiti specifici. Al livello ISCED 3, vengono impiegate per valutare e progettare sistemi di IA utilizzando la XAI per poter verificare adeguatezza, limiti e capacità di modelli e architetture.

Progettazione di sistemi di IA a livello di creazione: la XAI facilita il miglioramento continuo dei modelli fornendo informazioni dettagliate sul loro funzionamento, il che consente di identificare le aree di ottimizzazione, correggere potenziali errori e ridurre al minimo le distorsioni in ogni iterazione. Inquadrando questo livello attraverso l'ISCED, gli studenti di livello ISCED 1 acquisiscono una conoscenza di base del funzionamento dell'IA attraverso attività semplici e creative; al livello ISCED 2, iniziano a esplorare il miglioramento iterativo e la correzione delle distorsioni nei modelli di IA attraverso esperimenti guidati e, al livello ISCED 3, si confrontano in modo più approfondito con strumenti di IA, inclusa la spiegabilità, per analizzare criticamente, ottimizzare e ridurre i bias nella progettazione dei sistemi di IA.

Livelli ISCED 3-5 (istruzione e formazione professionale - VET)

La VET si basa su materie e professioni specifiche e utilizza l'IA in molti modi. L'intelligenza artificiale può essere utilizzata come strumento di insegnamento e apprendimento o come materia curricolare, e le materie devono evolversi per promuovere la spiegabilità. L'alfabetizzazione digitale critica è centrale nell'ambito dell'IA nella VET. In questo contesto, sono richieste le seguenti competenze per gli insegnanti in relazione alla XAI:

Collaborazione con l'IA: nella VET, la competenza nell'IA non si limita alla comprensione tecnica, ma comprende anche la capacità di collaborare con i sistemi di IA, interpretarne correttamente i risultati e utilizzare gli strumenti disponibili per migliorare il processo decisionale e la risoluzione dei problemi all'interno di uno specifico ambito professionale. Ad esempio, gli assistenti sanitari dovranno essere in grado di interagire con sistemi di triage basati sull'IA e di valutarne criticamente le indicazioni.

Competenze pratiche (Hands-on): L'alfabetizzazione all'intelligenza artificiale dovrebbe includere anche lo sviluppo di competenze pratiche. Per molti studenti della formazione professionale, ciò significa acquisire abilità tecniche utili a gestire, mantenere e risolvere problemi relativi a macchinari o software basati sull'IA, comprendendone i risultati e le risposte. Ad esempio, gli elettricisti potrebbero trovarsi a dover lavorare con sistemi domotici intelligenti dotati di XAI, ed è quindi fondamentale che siano in grado di interpretarli e gestirli in modo consapevole.

Uso etico degli strumenti di IA nell'industria: l'alfabetizzazione all'IA nella formazione professionale deve porre grande attenzione agli aspetti etici e alla sicurezza nell'uso delle tecnologie basate sull'intelligenza artificiale. In settori in cui l'IA svolge un ruolo cruciale nei processi decisionali — come la sanità, i trasporti o la produzione industriale — i lavoratori devono essere consapevoli delle implicazioni etiche associate alla professione, come il rischio di distorsioni negli algoritmi, l'impatto dell'automazione sui ruoli lavorativi e l'importanza della tutela della privacy e della sicurezza dei dati. Nel contesto della VET, ciò significa formare gli studenti affinché siano in grado di valutare in modo critico gli output e le decisioni generate dai sistemi di IA, in particolare in situazioni in cui è in gioco la sicurezza delle persone, come nella produzione automatizzata o nella diagnostica medica supportata dall'IA.

Apprendimento permanente: l'alfabetizzazione all'IA nella formazione professionale dovrebbe essere allineata anche con la crescente necessità di apprendimento permanente. Con l'evoluzione rapida delle tecnologie basate sull'IA, i lavoratori dovranno aggiornare costantemente le proprie competenze per restare competitivi. Promuovere l'alfabetizzazione all'IA nella VET significa quindi incoraggiare gli studenti ad assumersi la responsabilità del proprio apprendimento continuo, mantenendo e sviluppando nel tempo conoscenze e abilità in linea con le esigenze del settore, le tendenze e i progressi tecnologici. Poiché la XAI sarà una componente fondamentale di tutti gli strumenti basati sull'IA del futuro, è essenziale che venga inclusa anche nei programmi di aggiornamento e riqualificazione professionale.

Di conseguenza, i programmi di formazione professionale (VET) devono integrare l'alfabetizzazione all'IA nelle iniziative di aggiornamento e riqualificazione, supportando i lavoratori nella transizione verso nuovi ruoli che prevedono la supervisione di sistemi di IA, la gestione di processi integrati con l'IA o lo sviluppo di soluzioni basate sull'IA nei rispettivi settori. In questo modo si garantisce che i lavoratori restino competitivi e flessibili in un contesto in continua evoluzione tecnologica. Ad esempio, il progetto Erasmus+ *AI Pioneers* (Bekiaridis & Atwell, 2023), orientato al futuro, ha sviluppato una proposta di estensione del *Quadro europeo per le competenze digitali degli educatori: DigCompEdu (* Redecker & Punie, 2017) per insegnanti e formatori professionali in Europa,

che descrive in dettaglio le seguenti competenze per l'insegnamento e l'apprendimento con l'IA. I livelli sono coerenti con il framework DigCompEdu.

Livelli ISCED 6-8 (istruzione superiore)

Nel trattare le competenze legate alla XAI nell'istruzione superiore, è importante distinguere tra quelle richieste ai ricercatori e quelle necessarie agli insegnanti dei corsi di laurea tecnici e non tecnici. In ogni caso, si raccomanda che gli educatori a questo livello abbiano già acquisito le competenze di base previste per i livelli ISCED da 1 a 3. Le competenze richieste nell'istruzione superiore si caratterizzano per un approccio avanzato, autonomo e orientato alla ricerca, sia nell'insegnamento che nell'apprendimento. Inoltre, la spiegabilità assume un ruolo fondamentale nella ricerca, poiché è essenziale che i ricercatori siano in grado non solo di spiegare come utilizzano l'IA, ma anche di renderne trasparente il funzionamento.

Competenze dei ricercatori che utilizzano l'IA

Promuovere buone pratiche di scienza aperta e la trasparenza nella ricerca sull'IA: i ricercatori che utilizzano l'IA devono impegnarsi a rispettare i principi della scienza aperta e della trasparenza. Ciò implica l'adozione di pratiche che garantiscano la riproducibilità, l'accessibilità e la diffusione etica della ricerca. I ricercatori dovrebbero dare priorità alla condivisione di modelli, dataset e codice sorgente con licenze open source, al fine di permettere alla comunità scientifica di esaminare, validare e sviluppare ulteriormente i risultati. È inoltre fondamentale documentare in modo trasparente metodologie, ipotesi e limiti della ricerca, promuovendo così fiducia, inclusività e responsabilità. Questo approccio si allinea a quadri di riferimento internazionali come la Raccomandazione sull'Open Science dell'UNESCO. Infine, in coerenza con le linee guida dinamiche sull'uso responsabile dell'IA generativa nella ricerca della Commissione europea, i ricercatori dovrebbero adottare approcci responsabili e trasparenti nello sviluppo e nell'utilizzo degli strumenti di IA generativa.

Comprendere le tecniche XAI: è fondamentale saper scegliere quando applicare metodi di spiegabilità ante-hoc o post-hoc, in base al livello di interpretabilità del sistema di IA e ai requisiti della ricerca. La comprensione del ragionamento sotteso alle previsioni generate dai sistemi di intelligenza artificiale consente ai ricercatori di adottare decisioni più informate, fondate e consapevoli. Un ulteriore aspetto rilevante consiste nella capacità di distinguere tra le diverse dimensioni dell'XAI, come delineate nella Tabella 4. In particolare, la differenziazione tra spiegazioni di tipo locale — orientate all'interpretazione di singole previsioni — e spiegazioni di tipo globale — finalizzate a fornire una visione complessiva del comportamento del modello — risulta cruciale per l'identificazione di potenziali bias nei modelli di IA, quali il trattamento iniquo di gruppi demografici. Tale distinzione è fondamentale per promuovere principi di equità e trasparenza nei processi decisionali automatizzati.

Design centrato sulla persona: strettamente correlata alle competenze richieste nell'ambito dell'XAI è la padronanza dei principi del design centrato sulla persona, competenza fondamentale per i ricercatori che intendono integrare l'intelligenza artificiale nelle proprie attività. La trasparenza nella ricerca non può essere pienamente realizzata se il ricercatore non è in grado di modulare le spiegazioni in funzione dei diversi stakeholder, assicurando che esse risultino comprensibili, pertinenti e utili. Tale obiettivo è, tuttavia, raggiungibile solo a condizione che anche il modello di IA adottato sia intrinsecamente trasparente. I ricercatori in grado di articolare in modo efficace il funzionamento del modello sia a un pubblico tecnico sia a interlocutori non specialisti contribuiscono a rendere il proprio lavoro più accessibile, promuovendo una maggiore accettazione e comprensione dei risultati da parte di un pubblico eterogeneo.

Consapevolezza dell'impatto sociale e ambientale dell'IA: tale competenza consente di cogliere le implicazioni sistemiche derivanti dallo sviluppo e dall'implementazione dei sistemi di intelligenza artificiale, attraverso l'analisi critica di come tali tecnologie contribuiscano a modellare le norme sociali, sollevino dilemmi etici e comportino sfide di natura ambientale. I ricercatori che adottano una prospettiva critica nei confronti dell'IA — interrogandosi, ad esempio, su come essa possa privilegiare specifiche narrazioni culturali e/o linguistiche a discapito di altre — sviluppano pratiche di ricerca più responsabili, in grado di promuovere l'affidabilità, l'equità e la fiducia nei risultati prodotti.

Competenze richieste ai docenti di materie non tecniche

Alfabetizzazione critica alla XAI per un processo decisionale informato in ambiti specifici: i docenti universitari non tecnici dovrebbero essere in grado di valutare e interpretare criticamente il modo in cui strumenti e applicazioni basati sull'IA generano

risultati o raccomandazioni pertinenti alla propria area disciplinare. Ciò implica la consapevolezza dei meccanismi di base che rendono i sistemi di IA spiegabili, delle fonti comuni di bias e delle limitazioni derivanti da modelli opachi o proprietari. Fondamentale è anche la capacità di comunicare queste problematiche agli studenti, promuovendo un uso responsabile e basato su evidenze dell'IA nei diversi ambiti accademici e incoraggiando un atteggiamento critico che li accompagni nella loro futura carriera professionale.

Integrazione pedagogica dei principi XAI: oltre a saper interpretare criticamente i risultati generati dall'IA, i docenti non tecnici possono trarre beneficio dalla capacità di progettare o adattare attività didattiche che evidenzino i principi e le implicazioni della XAI nel proprio ambito disciplinare. Questo può includere, ad esempio, la creazione di compiti in cui gli studenti riflettano sul processo decisionale basato sull'IA in contesti reali, oppure attività in cui analizzano e confrontano risultati prodotti da sistemi di IA con ragionamenti guidati dall'uomo. Ciò comporta anche l'inserimento del tema della spiegabilità dell'IA (o della sua assenza) in discussioni più ampie su etica, equità e impatto sociale delle decisioni algoritmiche.

Competenze richieste ai docenti di materie tecniche

Conoscenza approfondita delle tecniche e degli algoritmi XAI: questa competenza permette di fornire agli studenti una comprensione approfondita delle varie tecniche e algoritmi XAI necessari per garantire la validità dei dati, dei modelli, dei processi, dei risultati e della portata delle spiegazioni. Gli educatori dovrebbero aiutare gli studenti ad acquisire una conoscenza approfondita dei principi, delle applicazioni e dei limiti di queste tecniche, consentendo loro di sviluppare sistemi di IA efficaci e trasparenti.

Integrare la XAI nella progettazione dei sistemi di IA: questa competenza serve per insegnare agli studenti come integrare la dimensione della spiegabilità a partire dalla fase iniziale di progettazione di un sistema di intelligenza artificiale. Gli educatori dovrebbero sottolineare l'importanza di progettare modelli di IA che non solo siano tecnicamente ben sviluppati, ma che seguano anche principi etici e forniscano spiegazioni chiare e comprensibili delle loro decisioni. In questo percorso, è fondamentale guidare gli studenti nella scelta di algoritmi e soluzioni progettuali che trovino il giusto equilibrio tra prestazioni elevate e trasparenza, così che sia gli utenti più esperti che quelli meno esperti possano capire la logica e i risultati prodotti dal sistema.

4.4 Competenze per le dimensioni chiave della XAI

L'integrazione di contenuti specifici relativi all'IA spiegabile all'interno dei programmi didattici è un nuovo scenario dell'alfabetizzazione all'intelligenza artificiale, che i responsabili delle politiche educative degli Stati membri dell'Unione Europea devono considerare. Tale integrazione favorirebbe una riflessione critica, sia da parte degli studenti sia degli insegnanti, sui processi decisionali adottati dai sistemi di IA, promuovendo un approccio più partecipativo, consapevole e riflessivo all'utilizzo delle tecnologie intelligenti. Inoltre, l'impiego dell'IA nei contesti didattici, supportato da strumenti basati su XAI, potrebbe rafforzare l'efficacia dell'apprendimento, grazie a percorsi personalizzati, a un maggiore coinvolgimento degli studenti e alla possibilità di interagire in tempo reale con i sistemi, comprendendone al contempo le logiche decisionali sottostanti. La XAI ha quindi un potenziale trasformativo nel rendere l'intelligenza artificiale comprensibile e utilizzabile per tutti gli studenti, indipendentemente dal loro livello di istruzione, e dovrebbe essere integrata nelle nuove forme di alfabetizzazione.

Le principali dimensioni della XAI, definite nella tabella 4, identificano quattro tipi principali di spiegazioni, che saranno presenti, in diversa misura, in tutti i sistemi di IA utilizzati nel prossimo futuro. Ogni utente di IA dovrebbe comprenderne le differenze e l'utilità, motivo per cui è necessario includerle nei percorsi di alfabetizzazione all'IA. Di conseguenza, gli insegnanti dovrebbero conoscere e comprendere le dimensioni della spiegabilità per poter formare i propri studenti. Inoltre, con tale conoscenza approfondita, gli educatori potrebbero promuovere quegli strumenti di IA che si adattano in modo più appropriato ai principi XAI e correlare le dimensioni chiave della XAI con gli obiettivi di apprendimento, le loro esigenze personali/istituzionali e quelle dei loro studenti.

Il quadro di competenze sull'IA dell'UNESCO (Cukurova & Miao, 2024) stabilisce le competenze richieste agli insegnanti per poter utilizzare gli strumenti tecnologici basati sull'IA nelle loro pratiche didattiche in modo sicuro, efficace ed etico. Poiché le competenze associate alle dimensioni chiave della XAI condividono tale obiettivo, è opportuno che siano sviluppate a partire dallo stesso quadro di riferimento. Tali competenze sono riportate nella tabella seguente.

Dimensione chiave	Competenze raccomandate
Ambito: globale vs locale	**Fondamenti e applicazioni dell'IA:** gli educatori devono comprendere sia le spiegazioni globali che quelle locali per valutare l'adeguatezza degli strumenti di IA e le loro implicazioni per contesti educativi specifici. Ciò include la valutazione del funzionamento degli strumenti di IA in vari scenari e la comprensione di comportamenti specifici dell'IA.
Approfondimento: completo vs selettivo	**Pedagogia dell'IA:** Gli educatori dovrebbero essere in grado di interpretare spiegazioni sia complete che selettive. Questo tipo di competenza consente loro non soltanto di fornire valutazioni dettagliate per revisioni approfondite dei sistemi, ma anche di offrire interpretazioni semplificate per un feedback immediato, elemento cruciale per un insegnamento e un apprendimento efficaci. **IA per lo sviluppo professionale:** Comprendere il livello di selettività nelle spiegazioni può supportare anche la crescita professionale degli insegnanti, aiutandoli ad adattare le proprie strategie didattiche in base agli spunti forniti dagli strumenti di intelligenza artificiale.

Alternative: spiegazioni contrastive vs non contrastivo	**Approccio incentrato sull'uomo ed etica dell'IA**: gli educatori devono essere in grado di fornire spiegazioni contrastive, aiutando studenti e stakeholder a comprendere le differenze tra diversi risultati. Questo implica una riflessione etica e un'assunzione di responsabilità nell'uso degli output dell'IA, affinché gli insegnanti possano giustificare le proprie decisioni e preservare l'integrità educativa. **Fondamenti e applicazioni dell'IA:** È inoltre fondamentale comprendere le spiegazioni non contrastive, poiché gli educatori devono saper identificare i fattori significativi che influenzano le decisioni dell'IA, anche senza dover necessariamente confrontare risultati differenti.
Flusso: condizionale vs correlazionale	**Pedagogia dell'IA:** gli educatori dovrebbero essere in grado di comunicare efficacemente le spiegazioni, utilizzando formati condizionali per garantire chiarezza e offrendo approfondimenti correlazionali per una comprensione più ampia. Questa competenza è fondamentale per integrare gli strumenti di IA nelle strategie pedagogiche. **IA per lo sviluppo professionale:** la capacità di trasmettere spiegazioni in modo chiaro ed efficace è essenziale per lo sviluppo professionale continuo degli educatori, permettendo loro di condividere conoscenze con i colleghi e di adattarsi alle nuove tecnologie di IA.

Tabella 13: Competenze degli educatori rispetto alle dimensioni chiave della XAI.

4.5 Applicazioni pratiche

Poiché le competenze richieste agli educatori variano in base ai diversi livelli ISCED, gli esempi di possibili applicazioni della XAI nell'educazione possono differire da un contesto all'altro. A partire da un'introduzione all'alfabetizzazione all'IA attraverso attività unplugged al livello ISCED 1, passando per lo sviluppo di competenze digitali avanzate legate all'IA ai livelli ISCED da 2 a 4, fino ad arrivare a un focus mirato nell'ambito dell'istruzione e formazione professionale (VET) o dell'istruzione superiore, i paragrafi che seguono propongono alcuni esempi di attività in cui la XAI svolge un ruolo chiave.

È importante sottolineare che la realizzazione concreta delle attività dipende da come – e da chi – vengono progettati i curricula nei vari Stati membri o dal fatto che tale progettazione rientri in progetti specifici, come è avvenuto per l'introduzione della programmazione e del pensiero computazionale.

Livello ISCED 1 (istruzione primaria)

Introduzione all'alfabetizzazione in materia di IA

A livello elementare, l'alfabetizzazione all'IA può iniziare con concetti semplici come il modo in cui le macchine apprendono e prendono decisioni. Strumenti didattici interattivi, come semplici giochi basati sull'IA o app di storytelling, potrebbero essere utilizzati per dimostrare come l'IA reagisce a diversi input (ad esempio, comandi vocali, riconoscimento facciale). Ad esempio, in classe si potrebbe usare un assistente virtuale basato sull'IA per aiutare gli alunni nella comprensione del testo. I docenti possono cogliere questa occasione per spiegare come l'IA interpreta le parole degli studenti e risponde di conseguenza, offrendo così una base introduttiva su come l'intelligenza artificiale elabora le informazioni.

Al livello ISCED 1, la tendenza è quella di cercare di ridurre al minimo il tempo trascorso davanti allo schermo.

Tuttavia, anche in questa fase iniziale è possibile mettere in pratica alcune attività di XAI. Potrebbe trattarsi di attività non digitali, che consentono ai giovani studenti di familiarizzare con il pensiero computazionale riflettendo sulla gamification e sull'uso di elementi quotidiani a loro familiari. Queste attività si svolgono nel contesto scolastico piuttosto che nella vita quotidiana dello studente, a casa. In questa fase, la XAI può essere integrata attraverso strumenti semplici e adatti ai bambini, in grado di spiegare perché l'IA propone determinati risultati.

Un insegnante della scuola primaria, ad esempio, potrebbe introdurre una applicazione basata sull'intelligenza artificiale per l'ortografia o la matematica, che non si limita a correggere le risposte, ma spiega anche la logica alla base degli errori. Al posto delle spiegazioni testuali, si potrebbero utilizzare spiegazioni vocali, codici colore o feedback sonori. Questo approccio favorisce lo sviluppo precoce del pensiero critico nei confronti dell'IA, aiutando gli studenti a comprendere che le macchine seguono regole e dati specifici, e incoraggiandoli a interrogarsi su tali processi.

Il progetto Data Science Fiction Scratch, accompagnato da un testo esplicativo (Tzampazi, n.d.), è stato ideato per esaminare la trasparenza nel punto di incontro tra pensiero critico, informatica e alfabetizzazione all'IA. Questo progetto funge sia da risorsa interattiva che da lezione di coding, sottolineando che i pregiudizi possono emergere anche senza l'IA.

— E figuriamoci quando entra in gioco l'IA, che è fondamentalmente modellata dall'input umano. Le competenze di base nella programmazione aiutano a demistificare questi processi, a sfatare i miti e ad aumentare la trasparenza dei sistemi. Sebbene molte piattaforme di machine learning di facile utilizzo e progetti rivolti ai bambini si concentrino sulle relazioni input-output e sulla trasparenza nella raccolta dei dati, spesso il funzionamento interno dei modelli non viene spiegato.

A differenza di un algoritmo di IA, questo progetto pone l'accento sugli aspetti algoritmici dell'elaborazione dei dati e del processo decisionale, offrendo un esempio concreto e una spiegazione sintetica del motivo per cui l'alfabetizzazione all'IA — parte essenziale dell'alfabetizzazione informatica — è strettamente interconnessa con la programmazione, i dati e l'alfabetizzazione matematica, quest'ultima costituendo una base fondamentale.

Alla radice di tutto ciò c'è alfabetizzazione critica: perché, come e chi può manipolare i dati? Questo rafforza la necessità di trasparenza, rendendo inaccettabili i modelli opachi — siano essi black box o locked box.

L'iniziativa AI Unplugged offre attività introduttive all'intelligenza artificiale che non richiedono un computer e sono adeguate a questo livello ISCED. Ad esempio, nel gioco Good-Monkey-Bad-Monkey, gli studenti creano un modello di classificazione utilizzando un albero decisionale. Analizzano un insieme di carte illustrate con scimmie che mostrano diverse espressioni e devono assegnarle a due categorie possibili: "buona" o "cattiva". Successivamente, costruiscono l'albero decisionale utilizzando le espressioni selezionate come caratteristiche per la creazione dei rami. L'obiettivo è testare il modello con nuove carte, rendendosi conto che l'algoritmo non è perfetto, poiché le categorie sono soggettive. Questo offre l'opportunità agli insegnanti di introdurre i concetti di accuratezza e di probabilità di successo. Nel contesto della XAI, gli insegnanti potrebbero adattare l'attività per includere una spiegazione della caratteristica utilizzata per determinare la scelta di un ramo piuttosto che un altro, stimolando così la riflessione sul processo decisionale dell'algoritmo.

Livello ISCED 2 (istruzione secondaria inferiore)

Sviluppare l'alfabetizzazione all'IA

Man mano che gli studenti progrediscono, l'alfabetizzazione all'IA può approfondire il modo in cui l'intelligenza artificiale viene utilizzata. Gli studenti delle scuole medie possono iniziare a conoscere le applicazioni dell'IA in vari settori come la sanità, i trasporti e l'intrattenimento. Gli insegnanti possono avviare discussioni su come l'IA impatta la vita degli studenti, ad esempio attraverso i social media, i sistemi di raccomandazione e i dispositivi intelligenti.

Gli studenti possono inoltre essere stimolati con la partecipazione a progetti in cui si indaga un sistema di IA (come un algoritmo di raccomandazione musicale), spiegandone il funzionamento. Con l'utilizzo di strumenti come il framework dell'UNESCO sull'etica dell'IA, gli studenti possono anche valutare le potenziali distorsioni generate da questi sistemi.

Il framework offre linee guida rigorose per lo sviluppo e l'applicazione etica dell'intelligenza artificiale, ponendo particolare enfasi sui principi fondamentali di equità, trasparenza e responsabilità. Attraverso l'impiego di tale strumento, gli insegnanti sono in grado di effettuare una valutazione critica delle piattaforme basate su sistemi di IA, quali strumenti di valutazione automatizzata o sistemi di apprendimento adattivo, al fine di individuare eventuali bias insiti nei dati, nella progettazione o nei processi decisionali. In particolare, è possibile analizzare i meccanismi decisionali degli strumenti predittivi relativi alle performance degli studenti, garantendo che nessuno studente sia svantaggiato ingiustamente a causa di variabili come il genere, l'etnia o il background socioeconomico. Inoltre, il framework promuove l'adozione di modelli di intelligenza artificiale maggiormente trasparenti, incentivando gli sviluppatori a fornire spiegazioni esaustive circa i processi sottostanti alle previsioni generate dai sistemi di IA. Questo processo consente agli insegnanti di agire come mediatori responsabili tra i sistemi di intelligenza artificiale e gli studenti, garantendo pratiche educative conformi ai principi etici e di equità.

L'utilizzo dell'intelligenza artificiale spiegabile nelle piattaforme educative personalizzate rappresenta senza dubbio un'opportunità significativa per studenti e insegnanti. Tuttavia, per selezionare la piattaforma più idonea al contesto didattico e alle caratteristiche dei discenti, è necessario che gli insegnanti acquisiscano le competenze specifiche delineate nel presente capitolo. I sistemi basati sull'intelligenza artificiale sono in grado di fornire un feedback dettagliato sui compiti, illustrando come il lavoro dello studente si posizioni rispetto a quello dei pari e motivando le raccomandazioni per il miglioramento. Tale approccio potrebbe stimolare negli studenti una valutazione critica dei sistemi stessi e della loro accuratezza.

Alcune idee per l'applicazione della XAI in classe

Per approfondire la competenza in XAI nelle scuole secondarie, gli educatori possono implementare diverse attività pratiche che combinano l'esplorazione tecnica con l'analisi etica. Ad esempio, in una classe di matematica o di informatica, gli studenti potrebbero costruire modelli di apprendimento supervisionato di base, come la regressione lineare o gli alberi decisionali, per comprendere come gli algoritmi "apprendono" dai dati di input. Sperimentando con diversi set di dati — come dati demografici o ambientali — gli studenti potranno poi osservare come la modifica delle variabili di input influisce sui risultati del modello e quindi capire come viene influenzato il processo di selezione dei dati.

Nelle scienze umane e sociali, gli insegnanti potrebbero accompagnare gli studenti nell'analisi di casi di studio relativi alle applicazioni dell'intelligenza artificiale in contesti reali, quali il *predictive policing*, dove gli algoritmi valutano il rischio sulla base di dati storici. Tale approccio consente agli studenti di esaminare come i bias presenti nei dati, ad esempio nei registri degli arresti, possano contribuire alla perpetuazione di disuguaglianze sociali, stimolando riflessioni sulle implicazioni etiche legate all'impiego dell'intelligenza artificiale. Inoltre, attraverso l'utilizzo di piattaforme digitali, gli studenti possono sviluppare modelli di IA che rappresentano visualmente le predizioni, consentendo loro di mettere alla prova le ipotesi formulate e di esplorare i confini decisionali.

Nel contesto dell'apprendimento dell'inglese come seconda lingua (ESL), gli studenti potrebbero utilizzare gli strumenti XAI per comprendere e valutare i modelli linguistici e le applicazioni di controllo grammaticale. Ad esempio, sperimentando strumenti grammaticali basati sull'intelligenza artificiale, gli studenti potranno osservare come l'applicazione suggerisca modifiche in base a specifiche regole linguistiche. Gli insegnanti potrebbero guidare gli studenti nell'analisi delle motivazioni alla base delle raccomandazioni dell'IA, stimolando una riflessione sulla sintassi, sulla scelta delle parole e sul contesto.

Promuovere una cultura in classe che incoraggi a mettere in discussione i risultati prodotti dall'IA è fondamentale. Gli insegnanti potrebbero organizzare sessioni di dibattito nelle quali gli studenti possano discutere sulle implicazioni etiche delle applicazioni di IA studiate, considerando aspetti quali la protezione dei dati, la trasparenza e la responsabilità. Integrando esercizi tecnici, casi di studio reali e riflessioni etiche, gli educatori preparano gli studenti non solo a comprendere i meccanismi dell'IA, ma anche a interagire in modo critico e responsabile con i sistemi di IA che incontreranno in futuro.

Livello ISCED 3 (istruzione secondaria superiore)

Approfondimento delle competenze in materia di IA

Nella scuola secondaria di secondo grado gli studenti potrebbero approfondire aspetti più avanzati dell'IA, includendo le sue implicazioni etiche e l'impatto sociale. Tali approfondimenti dovrebbero stimolare il pensiero critico e favorire discussioni sul ruolo dell'IA in ambiti quali l'applicazione della legge, la selezione del personale o l'assistenza sanitaria. Gli studenti possono fare riferimento all'AI Act, che sottolinea l'importanza della trasparenza e dell'equità nei sistemi di IA, per comprendere come le normative mirino a prevenire danni e garantire la responsabilità.

Un'attività interessante potrebbe consistere nell'analisi, da parte degli studenti, di casi concreti in cui l'impiego dell'intelligenza artificiale ha prodotto conseguenze indesiderate a causa di bias. Un esempio rilevante è rappresentato dalla tecnologia di riconoscimento facciale. Gli studenti potrebbero esaminare casi in cui gli algoritmi di riconoscimento facciale hanno manifestato bias, ad esempio identificando erroneamente persone appartenenti a determinati gruppi etnici o generi con maggiore frequenza. Tale attività metterebbe in luce l'importanza dell'equità e della responsabilità nella progettazione dei sistemi di IA.

Gli studenti potrebbero utilizzare piattaforme web open-source per la creazione e l'addestramento di modelli di machine learning di base. Attraverso l'impiego di tali strumenti, è possibile dimostrare concretamente come le modifiche nei dati di addestramento influenzino l'accuratezza del modello e l'insorgenza di bias. Queste piattaforme risultano particolarmente adatte ai contesti educativi in quanto semplificano concetti complessi legati all'intelligenza artificiale e offrono un riscontro visivo immediato. In un'ottica di preparazione degli studenti all'istruzione terziaria, gli insegnanti potrebbero proporre situazioni più vicine a scenari reali, affinché i requisiti relativi alla XAI risultino chiari e meglio delineati.

Esempio di applicazione della XAI

Per aiutare gli studenti a comprendere il concetto di distorsioni sistemiche dell'intelligenza artificiale, gli insegnanti possono guidarli in un semplice progetto che preveda la creazione, l'addestramento e il test di un modello di machine learning. Gli studenti imparano che i sistemi di IA "apprendono" dai dati di addestramento e che la qualità e l'equilibrio di tali dati influenzano direttamente l'accuratezza e l'equità del modello.

L'attività proposta può essere sviluppata in quattro fasi:

1. **Configurazione del modello:** gli studenti costruiscono un modello base di classificazione delle immagini (ad esempio, per identificare frutti o oggetti).
2. **Addestramento del modello:** gli studenti caricano ed etichettano immagini per diverse categorie (es. mele, banane, arance).
3. **Esperimenti sui dati:** gli studenti confrontano due scenari di addestramento — uno con dati bilanciati e uno con dati sbilanciati (ad esempio, 100 immagini di mele ma solo 10 di banane e arance).
4. **Test del modello:** gli studenti testano il modello con nuove immagini e osservano come l'utilizzo di set di dati bilanciati porti a predizioni più eque e accurate.

Un'altra applicazione della tecnologia dell'intelligenza artificiale per promuovere il pensiero critico e insegnare agli studenti delle scuole superiori l'importanza della trasparenza e della responsabilità dell'IA potrebbe consistere nello sviluppo di un progetto creativo che combini l'analisi letteraria con strumenti basati sull'IA.

In questa attività didattica, gli studenti sono invitati a utilizzare un generatore di arte basato sull'intelligenza artificiale per creare rappresentazioni visive di temi chiave, ambientazioni o simboli tratti da un'opera letteraria a loro scelta, come ad esempio *1984* di George Orwell. Il progetto favorisce l'apprendimento interdisciplinare, unendo letteratura inglese, arte e tecnologia, e incoraggia al contempo una riflessione critica sulle capacità e i limiti dell'intelligenza artificiale

Come nell'esempio precedente, relativo all'utilizzo di piattaforme web open source per creare e addestrare semplici modelli di machine learning, questo progetto è suddiviso in diverse fasi, ognuna delle quali ha un obiettivo preciso:

1. Preparazione

Analisi letteraria: gli studenti individuano i temi e i simboli principali presenti in 1984 (ad esempio, il Grande Fratello, il teleschermo) e formulano prompt descrittivi da utilizzare con il generatore di IA.

Introduzione all'IA: gli insegnanti spiegano come gli strumenti artistici basati sull'IA generano output visivi a partire dai prompt e introducono concetti come i dati di addestramento e l'emulazione dello stile.

g

2. Esecuzione

Creazione di arte con l'IA: gli studenti inseriscono prompt (ad esempio, "città distopica sotto sorveglianza") nel generatore di IA per visualizzare i temi. Creano più iterazioni per perfezionare i risultati.

Miglioramento operato dall'essere umano: gli studenti migliorano le immagini generate dall'IA utilizzando tecniche artistiche tradizionali o digitali, regolando dettagli, colori e composizione per rappresentare meglio l'atmosfera della storia e i simboli chiave.

3. Valutazione critica

Analisi dell'IA: gli studenti valutano quanto efficacemente l'IA sia riuscita a cogliere i temi di 1984, individuando eventuali elementi mancanti o sfumature trascurate.

Intelligenza artificiale vs creatività umana: si analizza in che misura l'intelligenza artificiale possa eguagliare la profondità emotiva espressa dagli artisti umani, riflettendo criticamente sulle differenze interpretative tra macchine ed esseri umani in relazione a metafore, contesti culturali e significati simbolici.

4. Discussione guidata dall'insegnante

Gli insegnanti guidano una discussione sui limiti creativi dell'intelligenza artificiale, evidenziando come essa si basi su dati di addestramento, pattern e algoritmi, ma sia priva della capacità di interpretare significati simbolici più profondi o livelli emotivi complessi. Tale riflessione può essere orientata sul ruolo insostituibile dell'essere umano nell'ambito artistico, sottolineando come l'intuizione e la sensibilità umana rappresentino elementi fondamentali nella creazione e attribuzione di significato — aspetti che le macchine non sono in grado di replicare.

In termini di risultati di apprendimento, gli studenti svilupperanno ulteriormente competenze di pensiero critico, in quanto saranno in grado di valutare la capacità dell'intelligenza artificiale di comprendere e rappresentare concetti propri dell'esperienza umana, acquisendo al contempo consapevolezza dell'importanza dell'apprendimento interdisciplinare per comprendere il funzionamento della creatività guidata dall'IA. Infine, questa attività pratica potrà favorire una riflessione sul delicato equilibrio tra il contributo umano e quello della macchina nel processo creativo.

Livelli ISCED 4-8 (istruzione post-secondaria non terziaria, istruzione terziaria di ciclo breve, laurea triennale o equivalente, laurea magistrale o equivalente, dottorato o equivalente)

Nell'istruzione superiore, è necessario operare una distinzione fra gli studenti dei corsi di laurea tecnici, che diventeranno sviluppatori di intelligenza artificiale o utenti esperti e quelli dei corsi di laurea non tecnici, che saranno utenti standard.

Competenze di livello avanzato in materia di IA

Nell'ambito dell'istruzione superiore, gli studenti iscritti a corsi di laurea di natura tecnica saranno esposti alle complessità teoriche e pratiche connesse allo sviluppo e all'implementazione dei sistemi di intelligenza artificiale. I programmi formativi potranno essere orientati alla progettazione, valutazione critica e distribuzione etica di tali sistemi. Un'attenzione particolare potrà essere rivolta all'analisi dei rischi e dei benefici derivanti dall'impiego dell'IA in ambiti ad alta rilevanza sociale, quali la medicina, il diritto e la gestione aziendale. Inoltre, sarà fondamentale promuovere lo studio delle modalità attraverso cui rendere i sistemi di IA trasparenti, spiegabili e responsabili, competenze imprescindibili per l'assunzione di ruoli professionali qualificati in questo settore.

Agli studenti universitari, in particolare nei corsi di informatica o di etica applicata, potrà essere richiesto di progettare e sviluppare propri modelli di IA, avvalendosi di strumenti di XAI, per esplicitare e motivare le decisioni adottate dai sistemi. Tali attività consentirebbero di acquisire competenze operative nella costruzione dell'IA e, contestualmente, di promuovere una cultura della trasparenza e della responsabilità nei confronti degli utenti finali.

Inoltre, in un'ottica di integrazione verticale del sistema educativo, gli studenti dell'istruzione superiore potrebbero contribuire allo sviluppo di strumenti XAI – quali chatbot o assistenti intelligenti – destinati alla didattica nella scuola primaria e secondaria. Ciò favorirebbe la creazione di un circuito virtuoso e continuo di trasmissione e co-costruzione della conoscenza, rafforzando il legame tra i diversi livelli del sistema formativo. L'adozione diffusa di approcci XAI all'interno dell'ecosistema educativo potrebbe infine generare soluzioni innovative e più efficaci per l'insegnamento e l'apprendimento.

Gli studenti dell'istruzione superiore potrebbero avvalersi dei sistemi di XAI per supportare le proprie attività di ricerca o lo svolgimento di elaborati accademici, ponendo tuttavia particolare attenzione alla comprensione critica delle implicazioni connesse all'affidamento all'IA. Ad esempio, qualora tali strumenti vengano utilizzati per l'analisi di ampi insiemi di dati o per l'assistenza nella redazione di relazioni scientifiche, l'integrazione delle tecniche XAI risulterebbe fondamentale per ricostruire e rendere trasparente il processo decisionale seguito dal sistema nell'elaborazione dei risultati. Tale approccio favorirebbe una maggiore consapevolezza rispetto all'eventuale presenza di distorsioni, approssimazioni o inesattezze nei risultati prodotti, promuovendo così un uso più responsabile e informato delle tecnologie basate sull'IA in ambito accademico.

Esempio di applicazione della XAI

In ambiti di ricerca scientifica intensiva come la sociologia o la data science, gli studenti possono utilizzare strumenti come SHAP o LIME per scomporre e analizzare il modo in cui i modelli di intelligenza artificiale elaborano e interpretano i dati sociali. Ciò consentirebbe loro di mettere in discussione le assunzioni sottostanti ai modelli e di valutare criticamente la validità delle conclusioni emerse, individuando eventuali elementi che richiedano un'ulteriore analisi o revisione.

Approfondimento delle competenze in materia di IA

Per gli studenti di corsi di laurea non tecnici, i requisiti di alfabetizzazione sono simili a quelli del livello ISCED 3, ma più specifici. Poiché diventeranno professionisti in settori specifici, tali studenti hanno bisogno di competenze per utilizzare correttamente gli strumenti di IA nel loro ambito professionale, comprenderne le caratteristiche tecniche e mantenere una visione critica delle risposte fornite.

Alcune idee per la messa in pratica

Gli studenti universitari di medicina dovrebbero acquisire competenze nell'utilizzo di strumenti di diagnostica per immagini basati sull'intelligenza artificiale, in quanto tali tecnologie saranno sempre più presenti nella loro futura pratica clinica. Prima di affrontare una formazione specialistica su questi strumenti, un'attività introduttiva orientata allo sviluppo delle competenze fondamentali potrebbe consistere nell'addestramento di una rete neurale artificiale attraverso un'applicazione semplificata per la classificazione delle immagini. Gli studenti sarebbero così invitati a raccogliere immagini mediche da fonti open access o da database specializzati e a sviluppare un modello in grado di prevedere la probabilità di presenza di determinate patologie o condizioni cliniche.
Nel corso di tale attività, gli studenti si confronteranno con le difficoltà legate al raggiungimento di alti livelli di accuratezza predittiva. Sarà loro richiesto di ottimizzare le prestazioni del modello e di fornire una riflessione critica sul tasso di successo ottenuto, motivandone le cause. Attraverso questa esperienza, essi matureranno una maggiore consapevolezza delle sfide tecniche e metodologiche che caratterizzano l'impiego dell'IA in contesti ad alta sensibilità come quello medico. Tale consapevolezza sarà importante per la selezione e per un uso critico, responsabile e informato degli strumenti di IA nella futura attività professionale.

Nei programmi di formazione degli insegnanti, è essenziale che i corsisti non solo acquisiscano competenze relative all'intelligenza artificiale e imparino a utilizzarla come sistema o strumento a supporto dei processi di insegnamento e apprendimento, ma sviluppino altresì una comprensione approfondita del ruolo della XAI in ambito educativo. Tale formazione dovrebbe includere la promozione delle capacità di pensiero critico necessarie per valutare e impiegare efficacemente gli strumenti basati sull'IA nella generazione di contenuti didattici, nonché la sensibilizzazione degli studenti rispetto alle potenziali distorsioni insite nei risultati prodotti dall'IA. Queste competenze risultano particolarmente rilevanti per i futuri insegnanti di lingue straniere, che devono essere consapevoli delle implicazioni culturali e linguistiche che possono emergere nei risultati generati da tali tecnologie.

Gli studenti iscritti a programmi di formazione per l'insegnamento delle lingue straniere potrebbero essere coinvolti nell'utilizzo di strumenti di IA per produrre tre output chiave ai fini della pianificazione didattica: un programma del corso, un piano delle lezioni e un'attività di apprendimento. Nell'ambito di tale attività, gli studenti potrebbero impiegare strumenti di IA per la generazione di materiali quali programmi di studio, valutazioni, compiti e domande di quiz. Successivamente, i lavori prodotti verrebbero sottoposti di nuovo agli strumenti di IA al fine di ricevere feedback e suggerimenti di miglioramento. Il docente ha il compito di stimolare gli studenti a testare e perfezionare i propri prompt, al fine di ottimizzare la qualità del feedback generato dall'IA. Inoltre, l'educatore facilita una riflessione critica sui rischi associati all'impiego dell'IA nella generazione di contenuti, enfatizzando l'importanza di comprendere come i risultati prodotti possano riflettere specifici contesti culturali o linguistici, come ad esempio le diverse varietà dialettali di una lingua.

Durante l'intero processo, gli studenti vengono incoraggiati ad adottare un approccio critico, volto all'esplorazione e alla valutazione dell'orientamento culturale e linguistico degli strumenti di IA utilizzati, nonché a discutere su come tali orientamenti possano influenzare l'adeguatezza dei risultati per differenti gruppi di studenti. Al termine dell'attività, è prevista la stesura di una riflessione basata su domande aperte, quali: in che modo gli studenti hanno completato i compiti assegnati; quali suggerimenti hanno prodotto i risultati migliori e le ragioni di tale efficacia; come lo strumento ha contribuito a migliorare il programma o il piano di lezione; quali attività suggerite dall'IA sono state adottate o scartate e per quali motivi; quali feedback sono stati ritenuti utili o meno e perché; e infine, in che misura è stato considerato il contesto culturale e linguistico dello strumento di IA e come ciò abbia influenzato i risultati, in particolare nel contesto dell'insegnamento delle lingue straniere.

Questo caso d'uso si riferisce alle competenze di XAI relative alla comprensione delle implicazioni culturali e linguistiche (conoscenze), nonché alla contestualizzazione e adattamento degli strumenti di IA per allinearli agli obiettivi di apprendimento e alle esigenze culturali o linguistiche degli studenti (abilità). Si sottolinea inoltre la responsabilità etica dei futuri educatori di mettere in discussione i sistemi di IA ed esigere maggiore trasparenza, inclusività ed equità da parte degli sviluppatori, oltre a riflettere criticamente su come tali sistemi possano dare priorità a determinate narrazioni culturali e/o linguistiche, sotto-rappresentandone altre, e a promuovere un approccio didattico più equilibrato e inclusivo (valori).

4.5 Raccomandazioni per gli stakeholder

La tabella seguente riassume le principali azioni che potrebbero essere intraprese dai diversi attori del sistema per integrare in modo adeguato la XAI nell'istruzione. Tali azioni completano quelle emerse dalle raccomandazioni politiche del workshop della comunità XAI dell'EDEH 2024, al fine di definire una roadmap formale a supporto dell'integrazione della XAI.

Stakeholder	Azioni chiave	Descrizione
Educatori	**Sviluppare una comprensione generale del funzionamento dei sistemi di IA e adottare un approccio critico.**	Prestare attenzione ai contenuti generati dall'IA che possono essere inaccurati o distorti e cercare di comprenderne i risultati.
	Garantire l'allineamento costruttivo degli obiettivi educativi con gli strumenti di IA.	Gli strumenti di IA dovrebbero supportare i risultati dell'apprendimento e allinearsi alle strategie di insegnamento e valutazione.
	Lavorare con l'IA.	Combinare i punti di forza dell'IA con il giudizio e l'esperienza personali per creare un equilibrio.
	Partecipare alla formazione professionale per migliorare le proprie competenze in materia di IA.	Tenersi aggiornati sugli sviluppi dell'IA e sulle sue applicazioni nelle pratiche didattiche.
Leader educativi	**Scegliere strumenti di IA che seguono i principi XAI.**	Assicurarsi che le applicazioni di IA soddisfino i requisiti di trasparenza e responsabilità e siano in linea con gli obiettivi istituzionali.
	Fornire opportunità di formazione per gli insegnanti e per il personale amministrativo.	Supportare gli educatori nell'orientare gli studenti sul funzionamento generale dell'IA.
	Dare priorità all'adozione di soluzioni di IA che supportino gli educatori e siano allineate con gli obiettivi pedagogici.	Seleziona strumenti che abbiano dimostrato di apportare valore nell'ambito educativo e che siano stati validati da esperti.
	Incoraggiare una comunicazione chiara sui processi decisionali dell'IA.	Garantire che i soggetti interessati comprendano le ragioni alla base delle raccomandazioni dell'IA.
	Adottare principi di progettazione incentrati sull'essere umano nell'integrazione dell'IA all'interno del proprio istituto.	Assicurarsi che i sistemi di IA siano utilizzati in modo da dare priorità al benessere, alle esigenze e agli obiettivi educativi degli studenti e degli educatori.
Policy maker	**Definire una serie di competenze a livello europeo.**	Sviluppare un quadro europeo per l'alfabetizzazione all'IA. Standardizzare l'inclusione della XAI nei programmi di studio di tutti gli Stati membri e promuovere una formazione approfondita degli educatori per garantire un'istruzione inclusiva in materia di IA.
	Fornire accesso alla potenza di elaborazione per gli istituti di istruzione.	Istituire hub regionali di IA dotati di infrastrutture informatiche condivise. Promuovere partnership con aziende tecnologiche per l'accesso sovvenzionato all'IA e finanziare la creazione di piattaforme di facile utilizzo con strumenti XAI integrati.
	Promuovere l'uso delle risorse educative open access (OER).	Investire nella creazione e nella traduzione delle OER e offrire formazione agli educatori sull'uso e la creazione delle OER. Garantire che queste risorse siano multilingue e culturalmente inclusive per soddisfare le esigenze di studenti con background diversi.
	Aumentare i finanziamenti per l'IA nell'istruzione.	Avviare fondi dedicati a supportare le infrastrutture, la formazione e lo sviluppo di curricoli sull'IA. Fornire sovvenzioni per progetti innovativi sull'IA e finanziare la ricerca sulle tecnologie XAI.

Tabella 14: Principali azioni richieste ai diversi stakeholder per integrare la XAI nell'istruzione.

4.6 Sintesi e considerazioni finali

Una volta che l'alfabetizzazione digitale sull'intelligenza artificiale sarà adeguatamente implementata, studenti ed educatori saranno in grado di interagire in modo più critico con i sistemi di IA, promuovendo così la XAI.

Nella pratica educative quotidiana:

- gli studenti acquisiranno la capacità di considerare l'intelligenza artificiale come uno strumento da interrogare criticamente. Sia che la utilizzino per attività di ricerca, progetti o processi di apprendimento, si approcceranno ad essa con una mentalità orientata alla ricerca di trasparenza e equità.
- gli educatori integreranno gli strumenti di IA nelle loro pratiche didattiche, con la consapevolezza di dover spiegare agli studenti il funzionamento interno di questi strumenti. Faciliteranno inoltre le discussioni sulle implicazioni etiche dell'IA, utilizzando la XAI per illustrare come l'IA prende le decisioni e perché è importante comprendere tale processo, nel contesto dell'istruzione come altrove.

In conclusione, l'integrazione dell'alfabetizzazione all'intelligenza artificiale in tutti i livelli di istruzione favorisce la formazione di una generazione non solo competente nell'utilizzo delle tecnologie IA, ma anche capace di esercitare un'analisi critica riguardo al ruolo sociale di tali tecnologie, di adottare decisioni informate in merito alle loro implicazioni etiche e di partecipare attivamente al loro sviluppo responsabile.

4. Conclusione

Nell'ambito dell'istruzione la XAI va oltre la semplice dimensione tecnica, configurandosi come un requisito imprescindibile per la promozione dell'autonomia individuale, della trasparenza e della fiducia in contesti educativi sempre più dipendenti dall'impiego di strumenti basati sull'intelligenza artificiale per finalità didattiche e di apprendimento. Come illustrato nel presente rapporto, la XAI assume un ruolo fondamentale nell'allineare i sistemi di IA ai principi educativi, alle disposizioni normative vigenti e agli obiettivi pedagogici perseguiti.

Questo rapporto ha approfondito il quadro normativo in continua evoluzione, analizzando le implicazioni derivanti dall'AI Act e dal GDPR, nonché le modalità con cui tali regolamentazioni interagiscono con le pratiche educative. I casi d'uso presentati in questa sede evidenziano la complessità insita nel garantire la conformità alle normative vigenti senza compromettere la fruibilità e l'efficacia degli strumenti di IA per i diversi stakeholder dell'educazione. Dagli insegnanti agli studenti, fino ai progettisti e ai policy maker, ciascun soggetto assume responsabilità e aspettative specifiche, che richiedono risposte attraverso spiegazioni di IA adeguatamente calibrate e di rilevanza operativa.

Inoltre, l'integrazione della XAI nei contesti educativi implica una revisione delle competenze digitali, con particolare attenzione all'alfabetizzazione sull'intelligenza artificiale. Gli educatori devono essere muniti non solo delle competenze tecniche necessarie all'uso degli strumenti IA, ma anche delle capacità critiche indispensabili per interpretarne i risultati e garantirne un impiego eticamente responsabile. Parallelamente, gli studenti necessitano di un adeguato supporto, volto a sviluppare una comprensione consapevole e critica delle decisioni generate dall'IA, favorendo così processi di autonomia e di apprendimento autoregolato.

In ultima analisi, la XAI non rappresenta un punto di arrivo, bensì un processo condiviso di co-creazione. Costruire sistemi di IA trasparenti, equi e centrati sulla persona nel contesto educativo implica lo sviluppo di tecnologie non solo valide dal punto di vista tecnico, ma anche contestualmente appropriate, conformi alla normativa e pedagogicamente efficaci. Ciò richiede una collaborazione continua tra discipline, settori e gruppi di stakeholder. Man mano che l'IA continua a evolversi devono evolversi anche gli sforzi collettivi per garantire che il suo impiego nell'istruzione supporti, e non sostituisca, il giudizio, i valori e la supervisione attiva dell'essere umano.