SINTESI

A. I.

RISK MANAGEMENT FRAMEWORK

DATA CREAZIONE: 27 Aprile 2023

NOME FILE: ALDO PEDICO - SINTESI AI RMF + PLAYBOOK E IR 8312.DOCX

*Nel documento "*TASSONOMIA E TERMINOLOGIA DEGLI ATTACCHI E DELLE ATTENUAZIONI NELLA INTELLIGENZA ARTIFICIALE*" (tratto dal* NIST AI 100-2e2023*) pubblicato dal sottoscritto, si descrive la manipolazione dei dati e gli effetti derivanti da tale azione sulla Privacy e sulla Cybersecurity.*

In questo documento, ho integrato e riassunto le pubblicazioni:

1. NIST AI 100-1 - AI RISK MANAGEMENT FRAMEWORK (RMF), [vedi anche NIST AI RMF 2nd Draft]
2. NIST AI RMF PLAYBOOK (https://pages.nist.gov/AIRMF/),
3. NIST IR 8312.

Nel loro insieme, queste pubblicazioni affrontano i rischi della PROGETTAZIONE*, dello* SVILUPPO*, dell'*UTILIZZO *e della* VALUTAZIONE *di prodotti, servizi e sistemi di IA; questi aspetti rientrano in un contesto organizzativo per la Gestione del Ciclo di Vita dell'IA, identificando:*

1) *i principali protagonisti dei processi di gestione;*
2) *le caratteristiche e le proprietà dei sistemi IA;*
3) *il nucleo centrale o core all'interno del quale sono descritte le funzioni che determinano le azioni suddivise in Categorie e Sottocategorie per la Gestione dei Rischi IA;*
4) *la descrizione dei* QUATTRO PRINCIPI per l'ESPLICAZIONE dell'IA *comprendenti le proprietà fondamentali per i sistemi.*

Suggerisco agli interessati all'argomento di prestare particolare attenzione, studiandone i dettagli, al NIST AI 100-1 - AI Risk Management Framework (RMF), *per i motivi descritti di seguito.*

A) spiega:

- ✓ *la motivazione per lo sviluppo e l'utilizzo del Framework;*
- ✓ *l'inquadramento del rischio e dell'affidabilità dell'IA; inoltre,*
- ✓ *include la descrizione dei profili e del loro utilizzo.*

In particolare, questo documento affronta i principi di AFFIDABILITÀ dell'IA.

Questo sistema di trattamento delle informazioni è soggetto a rischi di cibersicurezza, di privacy ed altri.

La gestione del rischio informatico dell'IA non è diversa dalla gestione del rischio per altri tipi di applicazioni informatiche.

I rischi per qualsiasi software o sistema basato sulle informazioni si applicano all'IA, comprese le apprensioni relative alla sicurezza informatica, alla privacy, alla sicurezza e all'infrastruttura.

Gli effetti dei sistemi di IA possono essere caratterizzati a lungo o a breve termine, ad alta o a bassa probabilità, sistemici o localizzati, ad alto o a basso impatto.

Il sistema dei rischi di IA può derivare: 1) dai dati utilizzati per addestrare il sistema stesso, 2) dagli obiettivi definiti al sistema stesso dall'uomo, 3) dall'uso del sistema di IA o dall'interazione di persone con il sistema IA.

I sistemi di IA sono di natura socio-tecnica, il che significa che sono un prodotto dei complessi fattori umani, organizzativi e tecnici coinvolti nella loro progettazione, sviluppo e utilizzo.

Molte delle caratteristiche affidabili dell'IA – come gestione dei pregiudizi, equità, interpretabilità e privacy – sono direttamente collegate alle dinamiche sociali e al comportamento umano.

B) descrive:

- ✓ *Ciclo di vita dell'IA,*
- ✓ *Chi fa cosa ovvero gli attori coinvolti e le loro competenze,*
- ✓ *Rischi e Affidabilità dell'IA,*
- ✓ *Core ovvero le azioni, suddivise per Funzioni, appartenenti all'intero Processo di Gestione del Rischio nell'IA,*
- ✓ *Quattro Principi d'Esplicazione dell'IA [NIST IR 8312].*

INDICE DEGLI ARGOMENTI

Titolo_____ Pag.

1 - CICLO DI VITA DELL'INTELLIGENZA ARTIFICIALE

NIST evidenzia l'importanza di Test, Valutazione, Verifica *e* Convalida (TEST EVALUATION VERIFICATION VALIDATION - TEVV) *durante tutto il ciclo di vita dell'IA e generalizza il contesto operativo di un sistema IA.*

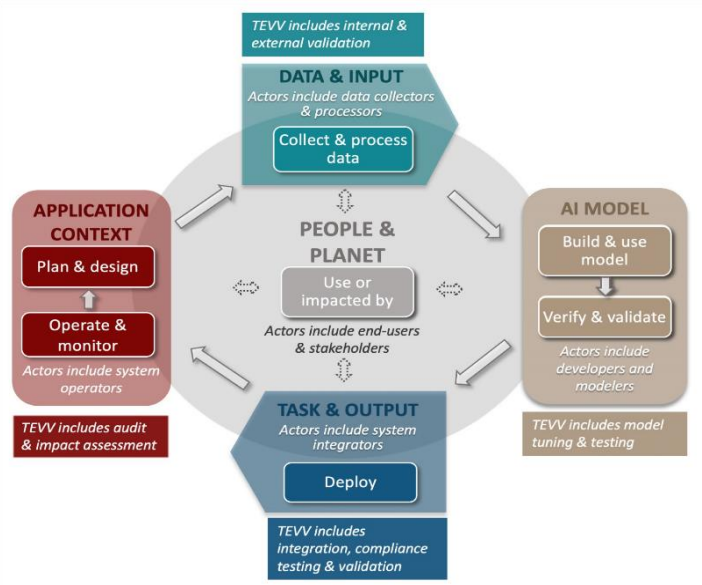


FIG. 1: CICLO DI VITA E DIMENSIONI DI UN SISTEMA DI IA

FIG. 2: ATTORI DELL'IA IN TUTTO IL CICLO DI VITA DELL'IA

Lifecycle	Activities	Representative Actors
Plan & design	Articulate and document the system's concept and objectives, underlying assumptions, context and requirements.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Collect & process data	Data collection & Processing: gather, validate, and clean data and document the metadata and characteristics of the dataset.	Data scientists, domain experts, socio-cultural analysts, human factors experts, data engineers, data providers, TEVV experts.
Build & use model	Create or select, train models or algorithms.	Modelers, model engineers, data scientists, developers, and domain experts. With consultation of socio-cultural analysts familiar with the application context, TEVV experts.
Verify & validate	Verify & validate, calibrate, and interpret model output.	
Deploy	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	System integrators, developers, systems/software engineers, domain experts, procurement experts, third-party suppliers with consultation of human factors experts, socio-cultural analysts, and governance experts, TEVV experts, end-users.
Operate & monitor	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Use or impacted by	Use system/technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.	End-users, affected individuals/communities, general public; policy makers, standards organizations, trade associations, advocacy groups, environmental groups, civil society organizations, researchers.

2 – INQUADRAMENTO DEL RISCHIO

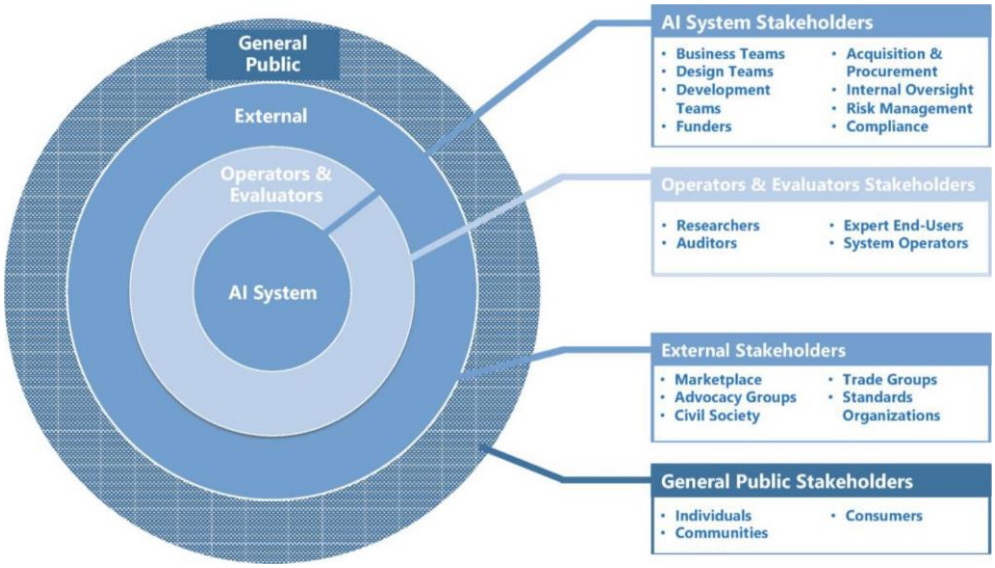
La gestione del rischio consiste nell'offrire un percorso per ridurre al minimo i potenziali impatti negativi sui sistemi, come le minacce alle libertà e ai diritti civili, oltre a indicare le opportunità per massimizzare gli impatti positivi.

Identificare, mitigare e ridurre *al minimo i rischi e i potenziali danni associati alle tecnologie di IA sono passi essenziali verso lo sviluppo di sistemi affidabili e il loro uso appropriato e responsabile.*

L'identificazione e la gestione dei rischi e degli impatti di AI, sia positivi che negativi, richiede un'ampia serie di prospettive e parti interessate.

FIG. 1: KEY STAKEHOLDER GROUPS ASSOCIATED WITH THE AI RMF

[NIST AI RMF 1st draft]



Come illustra la Figura 1, NIST ha identificato quattro gruppi di stakeholder come destinatari di questo Framework:

1. Stakeholder del sistema IA,
2. Operatori e Valutatori, Stakeholder esterni,
3. Pubblico in generale.

STAKEHOLDER SISTEMA IA

> *Sono coloro che hanno il maggior controllo e responsabilità sulla progettazione, sviluppo, implementazione e acquisizione di sistemi IA e l'implementazione delle pratiche di gestione del rischio IA.*

Questo gruppo comprende i principi adottati di IA RMF.
Possono includere individui o team all'interno o tra organizzazioni con responsabilità di commissionare, finanziare, acquistare, sviluppare o implementare un sistema IA: team aziendali, team di progettazione e sviluppo, team interni di gestione del rischio e team di conformità.

Le organizzazioni di piccole e medie dimensioni devono affrontare sfide diverse nell'attuazione di IA RMF rispetto alle grandi organizzazioni.

OPERATORI E VALUTATORI, STAKEHOLDER ESTERNI

*Forniscono il monitoraggio e il test formale/informale, la valutazione, la convalida e la verifica (*TEST, EVALUATION, VALIDATION AND VERIFICATION TEVV*) delle prestazioni del sistema, in relazione ai requisiti sia tecnici che socio-tecnici.*

Queste parti interessate, che includono organizzazioni che gestiscono o impiegano sistemi IA, utilizzano l'output per le decisioni o per valutare le proprie prestazioni.
Questo gruppo può includere utenti che interpretano o incorporano l'output dei sistemi IA in contesti con un alto potenziale di impatti negativi.
Potrebbero includere ricercatori del settore accademico, pubblico e privato; periti professionisti e revisori contabili; operatori di sistema; e utenti finali esperti.

PUBBLICO IN GENERALE

Forniscono norme o linee guida formali e/o quasi formali per specificare e affrontare i rischi di IA.

Esterni ai principali adottanti di IA RMF, possono includere gruppi commerciali, organizzazioni per lo sviluppo di standard, gruppi di difesa e organizzazioni della società civile.
Le loro azioni possono definire i confini per l'operazione (tecnici o legali) e bilanciare i valori e le priorità della società relativi alle libertà e ai diritti civili, all'economia e alla sicurezza.
È molto probabile che il pubblico subisca direttamente gli impatti positivi e negativi delle tecnologie IA.
Possono fornire la motivazione per le azioni intraprese dalle altre parti interessate e possono includere individui, comunità e consumatori nel contesto in cui un sistema IA è sviluppato o implementato.

2.1 - DIFFERENZA TRA I RISCHI IA E TRADIZIONALI

Analogamente al software tradizionale, i rischi derivanti dalla tecnologia basata sull'IA possono andare oltre all'azienda abbracciando le organizzazioni e impattando a livello sociale.

Le funzionalità di alcuni sistemi di IA presentano rischi seguenti: ad esempio, i modelli pre-addestrati e l'apprendimento di trasferimento possono far progredire la ricerca e aumentare l'accuratezza e la resilienza rispetto ad altri modelli e approcci.

I framework e le linee guida esistenti per la privacy, la sicurezza informatica e la sicurezza dei dati non sono in grado di:

1) *gestire adeguatamente il problema del* BIAS *nei sistemi di AI;*
2) *affrontare in modo completo i problemi di sicurezza relativi all'evasione, all'estrazione del modello, all'inferenza dell'appartenenza o ad altri attacchi di apprendimento automatico;*

3) *affrontare la complessa superficie di attacco dei sistemi di IA o altri abusi di sicurezza abilitati dai sistemi di IA; e*
4) *affrontare i rischi associati alle tecnologie di IA di terze parti, trasferire l'apprendimento e l'uso off-label, in cui i sistemi di IA possono essere addestrati per il processo decisionale al di fuori dei controlli di sicurezza di un'organizzazione o addestrati in un dominio e quindi "ottimizzati" per un altro.*

3 - AI RISKS AND TRUSTWORTHINESS

[fonte NIST AI RMF 1.0 draft]

Questo framework articola le seguenti caratteristiche di affidabilità e offre indicazioni per intraprenderle.

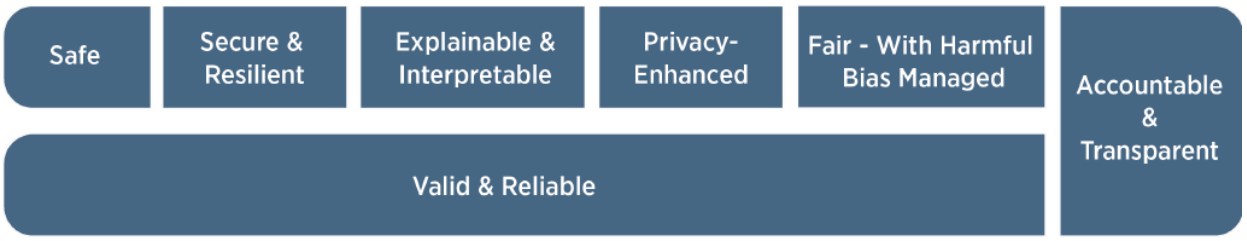
L'IA AFFIDABILE è:

1) *valida,*
2) *sicura,*
3) *equa,*
4) *resiliente,*
5) *responsabile e trasparente,*
6) *spiegabile e interpretabile,*

ed inoltre migliora la privacy.

FIGURE 4: AI TRUSTWORTHINESS CHARACTERISTICS

[fonte NIST AI 100-1 – AI RMF 1.0 final]



Queste caratteristiche sono legate al comportamento sociale e organizzativo umano, ai set di dati utilizzati dai sistemi di IA e alle decisioni prese da coloro che li costruiscono e alle interazioni con gli esseri umani che forniscono informazioni e supervisione di tali sistemi.

Sistemi di IA affidabili dovrebbero raggiungere un elevato grado di controllo sul rischio pur mantenendo un elevato livello di qualità delle prestazioni.

Il raggiungimento di questo difficile obiettivo richiede un approccio globale alla gestione del rischio, con compromessi tra le caratteristiche di affidabilità.

La decisione di commissionare o implementare un sistema di IA dovrebbe essere basata su una valutazione contestuale delle caratteristiche di affidabilità e dei relativi rischi, impatti, costi e benefici e informata da un ampio gruppo di parti interessate.

La Tabella 1 mappa la tassonomia IA RMF alla terminologia utilizzata dall'OCSE nella raccomandazione sull'IA, nella proposta di legge sull'IA dell'Unione europea (UE) e nell'ordine esecutivo degli Stati Uniti (EO) 13960.

TABELLA 1: MAPPATURA DELLA TASSONOMIA AI RMF AI DOCUMENTI DI POLICY AI

AI RMF	OECD AI Recommendation	EU AI Act (Proposed)	EO 13960
Valid and reliable	Robustness	Technical robustness	Purposeful and performance driven Accurate, reliable, and effective Regularly monitored
Safe	Safety	Safety	Safe
Fair and bias is managed	Human-centered values and fairness	Non-discrimination Diversity and fairness Data governance	Lawful and respectful of our Nation's values
Secure and resilient	Security	Security & resilience	Secure and resilient
Transparent and accountable	Transparency and responsible disclosure Accountability	Transparency Accountability Human agency and oversight	Transparent Accountable Lawful and respectful of our Nation's values Responsible and traceable Regularly monitored
Explainable and interpretable	Explainability		Understandable by subject matter experts, users, and others, as appropriate
Privacy-enhanced	Human values; Respect for human rights	Privacy Data governance	Lawful and respectful of our Nation's values

FATTORI UMANI

Poiché i sistemi di IA possono dare un senso alle informazioni in modo più rapido e coerente rispetto agli esseri umani, sono spesso implementati in contesti ad alto impatto come un modo per prendere decisioni più eque e imparziali rispetto al processo decisionale umano e per farlo in modo più efficiente. Una strategia comune per la gestione dei rischi in tali contesti è l'uso di un "in-the-loop" *umano (HITL).*

Pregiudizi possono essere indotti dagli attori dell'IA durante il ciclo di vita dell'IA tramite ipotesi, aspettative e decisioni durante le attività di modellazione.

Queste sfide sono esacerbate dall'opacità del sistema di IA e dalla conseguente mancanza di interpretabilità.

3.1 - VALIDO E AFFIDABILE

L'ACCURATEZZA e la ROBUSTEZZA sono fattori interdipendenti che contribuiscono alla validità e all'affidabilità dei sistemi.

L'implementazione di sistemi di IA imprecisi, inaffidabili o non generalizzabili ai dati al di là dei loro dati di addestramento (cioè non robusti) crea e aumenta i rischi e ne riduce l'affidabilità.

ACCURATEZZA – "**vicinanza dei risultati di osservazioni, calcoli o stime ai valori reali o ai valori accettati come veri**" *(Fonte: ISO/IEC TS 5723:2022) – dovrebbero affrontare sia gli aspetti computazionalmente-centrici (ad esempio, falsi positivi e falsi negativi) che gli aspetti di collaborazione uomo-IA.*

AFFIDABILITÀ – "**capacità di un articolo di funzionare come richiesto, senza guasti, per un determinato intervallo di tempo, in determinate condizioni**" *(Fonte: ISO/IEC TS 5723:2022) è un obiettivo per la correttezza complessiva del funzionamento del modello nelle condizioni di utilizzo previsto e in un determinato periodo di tempo, per includere l'intera durata del sistema.*

ROBUSTEZZA – "**capacità di un sistema di IA di mantenere il suo livello di prestazioni in una varietà di circostanze**" *(Fonte: ISO / IEC TS 5723: 2022) – è un obiettivo per una funzionalità di sistema appropriata in un ampio insieme di condizioni e circostanze, compresi gli usi di sistemi di IA non inizialmente previsti.*

3.2 – EQUITÀ (FAIR) CON LA GESTIONE DEI PREGIUDIZI (BIAS) DANNOSI

[fonte NIST AI 100-1 - AI RMF 1.0 final]

BIAS: *i* BIAS, *o meglio* BIAS *cognitivi, sono delle distorsioni che le persone attuano nelle valutazioni di fatti e avvenimenti. Tali distorsioni ci spingono a ricreare una propria visione soggettiva che non corrisponde fedelmente alla realtà. In sintesi, i* BIAS *cognitivi rappresentano il modo con cui il nostro cervello distorce di fatto la realtà.*

L'equità nell'IA genera apprensioni afferenti all'uguaglianza affrontando questioni come pregiudizi e discriminazioni.

Gli standard di equità possono essere complessi e difficili da definire perché le percezioni di equità differiscono tra le culture e possono cambiare a seconda dell'applicazione.

I sistemi in cui i pregiudizi sono mitigati non sono necessariamente equi. Ad esempio, i sistemi in cui le previsioni sono in qualche modo bilanciate tra i gruppi demografici possono ancora essere inaccessibili alle persone con disabilità o colpite dal divario digitale.

CATEGORIE DI PREGIUDIZI

NIST ha identificato tre principali CATEGORIE DI PREGIUDIZI *dell'IA da considerare e gestire:*

1) SISTEMICO,
2) COMPUTAZIONALE,
3) UMANO,

i quali possono verificarsi in assenza di PREGIUDIZIO, PARZIALITÀ *o* INTENTO DISCRIMINATORIO.

- ✓ *Il* PREGIUDIZIO SISTEMICO *può essere presente nei set di dati dell'IA, nelle norme, nelle pratiche e nei processi organizzativi in tutto il ciclo di vita dell'IA e nella società più ampia che utilizza i sistemi di IA.*
- ✓ *La* DISTORSIONE COMPUTAZIONALE *può essere presente nei set di dati e nei processi algoritmici dell'IA e spesso deriva da errori sistematici dovuti a campioni non rappresentativi.*
- ✓ *I pregiudizi* UMANI*:*
  - ➢ *si riferiscono al modo in cui un individuo o un gruppo percepisce le informazioni del sistema IA per prendere una decisione o compilare le informazioni mancanti;*
  - ➢ *sono onnipresenti nei processi decisionali durante il ciclo di vita dell'IA e l'uso del sistema;*
  - ➢ *sono impliciti, quindi aumentare la consapevolezza non assicura il controllo o il miglioramento.*

3.3 - SICURO E RESILIENTE

SICUREZZA *e* RESILIENZA *sono caratteristiche correlate ma distinte.*

Mentre la RESILIENZA *è la capacità di tornare alla normale funzione dopo un attacco, la* SICUREZZA *include la* RESILIENZA *ma comprende anche protocolli per proteggere dagli attacchi oppure evitarli.*

La RESILIENZA *ha una certa relazione con la* ROBUSTEZZA*, tranne per il fatto che va oltre la provenienza dei dati per comprendere l'uso imprevisto o contraddittorio del modello o dei dati.*

Altri problemi di SICUREZZA *comuni riguardano l'*AVVELENAMENTO *dei dati e l'*ESFILTRAZIONE *di modelli, dei dati di addestramento o di altra proprietà intellettuale attraverso gli endpoint del sistema di IA.*

3.4 - TRASPARENZA E RESPONSABILITÀ

La TRASPARENZA *riflette la misura in cui le informazioni sono disponibili per gli individui su un sistema di IA, se interagiscono – o addirittura sono consapevoli di interagire – con tale sistema.*

Il suo ambito spazia dalle decisioni di progettazione e dai dati di training al training del modello, alla struttura del modello, al caso d'uso previsto e a come e quando sono state prese le decisioni della distribuzione o dell'utente finale e da chi.

La RESPONSABILITÀ *si riferisce alle aspettative nel caso in cui si realizzi un risultato rischioso.*

La relazione tra RISCHIO *e* RESPONSABILITÀ *associata all'IA e ai sistemi tecnologici differisce più ampiamente tra contesti culturali, legali, settoriali e sociali.*

3.5 - ESPLICABILE E INTERPRETABILE

*L'*ESPLICAZIONE *si riferisce a una rappresentazione dei meccanismi alla base del funzionamento di un algoritmo.*

*L'*INTERPRETABILITÀ *si riferisce al significato dell'output dei sistemi di IA nel contesto del suo scopo funzionale progettato.*

Il rischio derivante dalla mancanza di SPIEGABILITÀ *può essere gestito da descrizioni di come funzionano i modelli su misura per le differenze individuali come il livello di conoscenza e abilità dell'utente.*

*I rischi per l'*INTERPRETABILITÀ *possono spesso essere affrontati comunicando una descrizione del motivo per cui un sistema di IA ha effettuato una particolare previsione o raccomandazione. (Vedi NISTIR 8312, "Quattro principi di IA spiegabile" e NISTIR 8367, "Fondamenti psicologici di spiegabilità e interpretabilità nell'artificiale Intelligenza").*

3.6 – PRIVACY ENHANCED

La PRIVACY *si riferisce generalmente alle norme e alle pratiche che aiutano a salvaguardare l'autonomia, l'identità e la dignità umana.*

*Queste norme e pratiche in genere riguardano la libertà dall'*INTRUSIONE, *la* LIMITAZIONE *dell'* OSSERVAZIONE, *della* DIVULGAZIONE *o al* CONTROLLO *di aspetti della loro identità (ad esempio, corpo, dati, reputazione). (Vedi Il NIST Privacy Framework: uno strumento per migliorare la privacy attraverso la gestione del rischio aziendale.)*

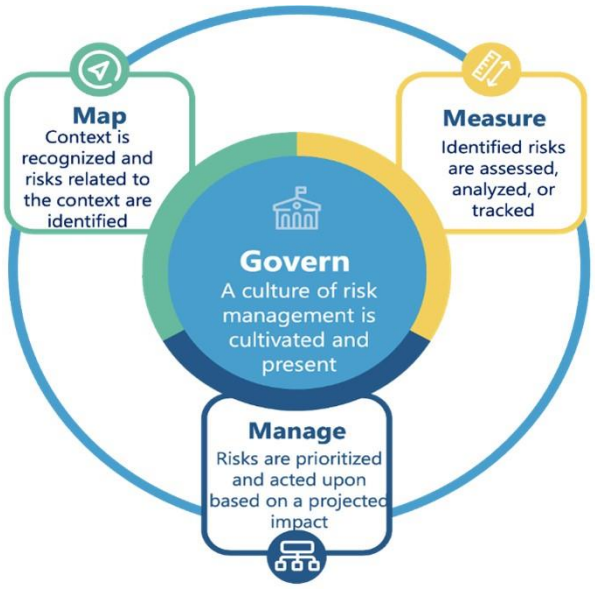
*I valori della privacy come l'*ANONIMATO, *la* RISERVATEZZA *ed il* CONTROLLO *in generale dovrebbero guidare le scelte per la* PROGETTAZIONE, *lo* SVILUPPO *e la* DISTRIBUZIONE *del sistema di IA.*

Dal punto di vista delle POLITICHE, **i rischi legati alla privacy possono sovrapporsi** *a* SICUREZZA, PREGIUDIZI *e* TRASPARENZA.

4 - CORE

[NIST AI RMF 1.0 finale]

FIG. 5: FUNZIONI PER LA GESTIONE DEL RISCHIO NELL'IA



IA RMF CORE fornisce i risultati e le azioni che consentono il dialogo, la comprensione e le attività per gestire i rischi nell'IA.

Come illustrato nella Figura 5, il CORE *è composto da quattro funzioni:*

1) MAP,
2) MEASURE,
3) MANAGE,
4) GOVERN.

Ognuna di queste funzioni di alto livello è suddivisa in CATEGORIE *e* SOTTOCATEGORIE.

Le CATEGORIE *e le* SOTTOCATEGORIE *sono suddivise in* RISULTATI *e* AZIONI *specifiche.*

Le AZIONI *non costituiscono una lista di controllo, né sono necessariamente un insieme ordinato di passaggi.*

Alcune organizzazioni possono scegliere tra le CATEGORIE e le SOTTOCATEGORIE; altri vorranno e avranno la capacità di applicarle tutte.

Supponendo che sia in atto una struttura di governance, le funzioni possono essere eseguite in qualsiasi ordine durante il ciclo di vita dell'IA.

Il processo dovrebbe essere iterativo, con riferimenti incrociati tra le funzioni se necessario.

Allo stesso modo, ci sono CATEGORIE e SOTTOCATEGORIE con elementi che si applicano a più funzioni o che devono verificarsi prima di determinate decisioni di sottocategoria.

Una risorsa complementare online all'IA RMF, denominata NIST IA RMF Playbook*, è disponibile per aiutare le organizzazioni a navigare nell'IA RMF e raggiungere i risultati attraverso azioni tattiche suggerite che possono applicare all'interno dei propri contesti.*

Insieme all'IA RMF, *il* NIST Playbook *farà parte del prossimo* "TRUSTWORTHY AND RESPONSIBLE AI RESOURCE CENTER".

GOVERN

La funzione GOVERN *coltiva e implementa una cultura della gestione del rischio all'interno delle organizzazioni che sviluppano, distribuiscono o acquisiscono sistemi di IA.*

GOVERN *è una funzione trasversale che è infusa in tutto ed informa le altre funzioni del processo.*

La governance:

1) *È progettata per garantire che i rischi e i potenziali impatti siano identificati, misurati e gestiti in modo efficace e coerente.*
2) *È progettata per fornire una struttura attraverso la quale le funzioni di gestione del rischio dell'IA possano allinearsi con le politiche organizzative e le priorità strategiche, indipendentemente dal fatto che siano correlate o meno ai sistemi di IA.*
3) *Si concentra sugli aspetti tecnici della progettazione e dello sviluppo del sistema di IA, nonché sulle pratiche e le competenze organizzative che influenzano direttamente le persone coinvolte nella formazione, nella distribuzione e nel monitoraggio di tali sistemi.*
4) *Dovrebbe riguardare le catene di approvvigionamento, compresi i sistemi e i dati software o hardware di terze parti, nonché i sistemi di IA sviluppati internamente.*

GOVERN *è una funzione trasversale che viene infusa in tutta la gestione del rischio IA e informa le altre funzioni del processo.*

Gli aspetti di GOVERN, *in particolare quelli relativi alla conformità o alla valutazione, dovrebbero essere integrati in ciascuna delle altre funzioni.*

L'attenzione alla governance è un requisito continuo e intrinseco per un'efficace gestione del rischio di IA nel corso della vita di un sistema di IA e della gerarchia dell'organizzazione.

La leadership definisce il tono per la gestione del rischio all'interno di un'organizzazione e, con essa, la cultura organizzativa.

I responsabili determinano le politiche generali che si allineano con la missione, gli obiettivi, i valori e la propensione al rischio dell'organizzazione, nonché con la sua cultura.

Le pratiche relative alla gestione dei rischi dell'IA sono descritte nel Playbook NIST AI RMF.

Nella tabella 2 sono elencate le categorie e le sottocategorie della funzione GOVERN.

TABLE 2: CATEGORIES AND SUBCATEGORIES FOR THE GOVERN FUNCTION

CATEGORIA	SOTTOCATEGORIA
Funzione **GOVERN**: una cultura della gestione del rischio è coltivata e presente	
GOVERN 1: Le politiche, i processi, le procedure e le pratiche in tutta l'organizzazione relative alla mappatura, alla misurazione e alla gestione dei rischi dell'IA sono in atto, trasparenti e implementate in modo efficace.	*GOVERN 1.1: I requisiti legali e normativi che coinvolgono l'IA sono compresi, gestiti e documentati.*
	GOVERN 1.2: Le caratteristiche dell'IA affidabile sono integrate nelle politiche, nei processi e nelle procedure organizzative.
	GOVERN 1.3: Il processo di gestione del rischio e i suoi risultati sono stabiliti attraverso meccanismi trasparenti e vengono misurati tutti i rischi significativi determinati.

	GOVERN 1.4: *Sono previsti il monitoraggio continuo e la revisione periodica del processo di gestione del rischio e dei suoi esiti, con ruoli e responsabilità organizzative chiaramente definiti.*
GOVERN 2: *Responsabilità* *le strutture sono in atto in modo che i team e gli individui appropriati siano responsabilizzati, responsabili e formati per la mappatura, la misurazione e la gestione dei rischi dell'IA.*	GOVERN 2.1: *I ruoli e le responsabilità e le linee di comunicazione relative alla mappatura, misurazione e gestione dei rischi dell'IA sono documentati e sono chiari agli individui e ai team in tutta l'organizzazione.*
	GOVERN 2.2: *Il personale e i partner dell'organizzazione ricevono una formazione sulla gestione del rischio IA per consentire loro di svolgere i propri compiti e responsabilità in linea con le politiche, le procedure e gli accordi correlati.*
	GOVERN 2.3: *La leadership esecutiva dell'organizzazione considera le decisioni sui rischi associati allo sviluppo e all'implementazione del sistema di IA come loro responsabilità.*
GOVERN 3: *Forza lavoro* *i processi di diversità, equità, inclusione e accessibilità hanno la priorità nella mappatura, misurazione e gestione dei rischi dell'IA durante l'intero ciclo di vita.*	GOVERN 3.1: *Il processo decisionale relativo alla mappatura, misurazione e gestione dei rischi dell'IA durante l'intero ciclo di vita è informato da un team demograficamente e disciplinarmente diversificato, incluso personale interno ed esterno. In particolare, i team che sono direttamente coinvolti nell'identificazione delle considerazioni e dei rischi di progettazione includono una diversità di esperienze, competenze e background per garantire che i sistemi di IA soddisfino i requisiti al di là di un ristretto sottoinsieme di utenti.*
GOVERN 4: *I team organizzativi sono impegnati in una cultura che considera e comunica il rischio.*	GOVERN 4.1: *Le pratiche organizzative sono in atto per promuovere un pensiero critico e una mentalità incentrata sulla sicurezza nella progettazione, nello sviluppo e nell'implementazione di sistemi di IA per ridurre al minimo gli impatti negativi.*
	GOVERN 4.2: *I team organizzativi documentano i rischi e gli impatti della tecnologia che progettano, sviluppano o distribuiscono e comunicano gli impatti in modo più ampio.*
	GOVERN 4.3: *Le pratiche organizzative sono in atto per consentire test, identificazione di incidenti e condivisione delle informazioni.*
GOVERN 5: *I processi sono in atto per un solido coinvolgimento degli stakeholder.*	GOVERN 5.1: *Le politiche e le pratiche organizzative sono in atto per raccogliere, considerare, dare priorità e integrare il feedback degli stakeholder esterni in merito ai potenziali impatti individuali e sociali relativi ai rischi dell'IA.*
	GOVERN 5.2: *Vengono istituiti meccanismi per consentire agli attori dell'IA di incorporare regolarmente il feedback delle parti interessate nella progettazione e nell'implementazione del sistema.*
GOVERN 6: *Sono in atto politiche e procedure per affrontare i rischi di IA derivanti da software e dati di terze parti e altri problemi della catena di approvvigionamento.*	GOVERN 6.1: *Sono in atto politiche e procedure che affrontano i rischi associati a entità di terze parti.*
	GOVERN 6.2: *Sono in atto processi di emergenza per gestire guasti o incidenti in dati di terze parti o sistemi di IA ritenuti ad alto rischio.*

MAP

La funzione MAP *stabilisce il contesto per inquadrare i rischi relativi a un sistema di IA.*

Le informazioni raccolte durante lo svolgimento di questa funzione consentono la prevenzione dei rischi e informano le decisioni per processi come la gestione del modello e una decisione iniziale sull'adeguatezza o sulla necessità di una soluzione di IA.

I risultati nella funzione MAP *sono la base per le funzioni* MEASURE *e* MANAGE.

La raccolta di ampie prospettive può aiutare le organizzazioni a prevenire in modo proattivo i rischi e a sviluppare sistemi di IA più affidabili:

- ✓ *migliorare la loro capacità di comprensione dei contesti;*
- ✓ *verificare le loro ipotesi sul contesto di utilizzo;*
- ✓ *consentire il riconoscimento di quando i sistemi non sono funzionali all'interno o all'esterno del contesto previsto;*
- ✓ *identificare gli usi positivi e benefici dei loro sistemi di IA esistenti e dei nuovi mercati;*
- ✓ *migliorare la comprensione delle limitazioni in processi come lo sviluppo di proxy;*
- ✓ *identificare i vincoli nelle applicazioni del mondo reale che possono portare a impatti negativi.*

Completata la funzione MAP, *gli utenti dovrebbero avere una conoscenza contestuale sufficiente degli impatti del sistema di IA per informare una decisione go/no-go sull'opportunità di progettare, sviluppare o distribuire un sistema di IA basato su una valutazione degli impatti.*

Se viene presa una decisione per procedere, le organizzazioni dovrebbero utilizzare le funzioni MEASURE *e* MANAGE *per assistere negli sforzi di gestione del rischio IA, utilizzando politiche e procedure messe in atto nella funzione* GOVERN.

Le pratiche relative alla mappatura dei rischi dell'IA sono descritte nel Playbook NIST AI RMF.

La tabella 3 elenca le categorie e le sottocategorie della funzione MAP.

TABLE 3: CATEGORIES AND SUBCATEGORIES FOR THE MAP FUNCTION

CATEGORIA	SOTTOCATEGORIA
Funzione MAP: il contesto viene riconosciuto e vengono identificati i rischi relativi al contesto	
MAP 1: Il contesto è stabilito e compreso	*MAP 1.1: Scopo previsto, impostazioni prospettiche in cui verrà implementato il sistema di IA, il set specifico o i tipi di utenti insieme alle loro aspettative e gli impatti dell'uso del sistema sono compresi e documentati. Le ipotesi e le relative limitazioni sullo scopo e l'utilizzo del sistema di IA sono enumerate, documentate e collegate alle considerazioni TEVV e alle metriche di sistema.*
	MAP 1.2: Gli attori, le competenze, le abilità e le capacità interdisciplinari dell'IA per stabilire il contesto riflettono la diversità demografica e l'ampia competenza nel dominio e nell'esperienza dell'utente e la loro partecipazione è documentata. *Le opportunità di collaborazione interdisciplinare sono prioritarie.*
	MAP 1.3: Il valore aziendale o il contesto dell'uso aziendale è stato chiaramente definito o, nel caso della valutazione dei sistemi di IA esistenti, rivalutato.
	MAP 1.4: La missione dell'organizzazione e gli obiettivi rilevanti per la tecnologia IA sono compresi.
	MAP 1.5: Vengono determinate le tolleranze al rischio organizzativo.
	MAP 1.6: Le pratiche e il personale per le attività di progettazione consentono un impegno regolare con le parti interessate e integrano feedback fruibili degli utenti e della comunità sugli impatti negativi imprevisti.
	MAP 1.7: I requisiti di sistema (ad esempio, "il sistema deve rispettare la privacy dei suoi utenti") sono suscitati e compresi dalle parti interessate. Le decisioni di progettazione tengono conto delle implicazioni socio-tecniche per affrontare i rischi dell'IA.

MAP 2: Viene eseguita la classificazione del sistema di IA	*MAP 2.1: Viene definito il compito specifico e i metodi utilizzati per implementare l'attività che il sistema di IA supporterà (ad esempio, classificatori, modelli generativi, raccomandatori).*
	MAPPA 2.2: Le informazioni sono documentate sui limiti di conoscenza del sistema e su come l'output sarà utilizzato e supervisionato dagli esseri umani.
	MAP 2.3: L'integrità scientifica e le considerazioni TEVV sono identificate e documentate, comprese quelle relative alla progettazione sperimentale, alla raccolta e alla selezione dei dati (ad esempio, disponibilità, rappresentatività, idoneità) e alla convalida del costrutto.
MAP 3: vengono comprese le funzionalità di IA, l'utilizzo mirato, gli obiettivi e i benefici e i costi previsti rispetto allo status quo	*MAP 3.1: I vantaggi delle funzionalità e delle prestazioni del sistema previste vengono esaminati e documentati.*
	MAP 3.2: I costi potenziali, compresi i costi non monetari, che derivano da errori attesi o realizzati o dalle prestazioni del sistema vengono esaminati e documentati.
	MAP 3.3: L'ambito dell'applicazione mirata viene specificato, ristretto e documentato in base al contesto stabilito e alla classificazione del sistema di IA.
MAP 4: i rischi e i benefici sono mappati per software e dati di terze parti	*MAP 4.1: Gli approcci per la mappatura dei rischi tecnologici di terze parti sono in atto e documentati.*
	MAP 4.2: I controlli interni dei rischi per i rischi tecnologici di terze parti sono in atto e documentati.
MAP 5: Vengono valutati gli impatti su individui, gruppi , comunità, organizzazioni e società	*MAP 5.1: I potenziali impatti positivi e negativi su individui, gruppi, comunità, organizzazioni e società sono regolarmente identificati e documentati.*
	MAP 5.2: La probabilità e l'entità di ciascun impatto identificato in base all'uso previsto, agli usi passati dei sistemi di IA in contesti simili, alle segnalazioni di incidenti pubblici, al feedback delle parti interessate o ad altri dati sono identificati e documentati.
	MAP 5.3: Le valutazioni dei benefici rispetto agli impatti si basano su analisi di impatto, entità e probabilità di rischio.

MEASURE

La funzione MEASURE *utilizza strumenti, tecniche e metodologie quantitative, qualitative o a metodo misto per analizzare, valutare, confrontare e monitorare il rischio dell'IA e gli impatti correlati.*

Utilizza le conoscenze relative ai rischi di IA identificati nella funzione MAP *e informa la funzione* MANAGE.

La misurazione dei rischi dell'IA include il monitoraggio delle metriche per caratteristiche di affidabilità, di impatto sociale e di configurazioni uomo-IA.

I processi sviluppati o adottati nella funzione MEASURE *dovrebbero includere rigorose metodologie di test del software e di valutazione delle prestazioni che includano misure associate di* INCERTEZZA, *confronti con gli* INDICI DI RIFERIMENTO DELLE PRESTAZIONI *e* RELAZIONI *formalizzate e* DOCUMENTAZIONE *dei risultati.*

Dopo aver completato la funzione MEASURE, *i processi* TEVV, *tra cui metriche, metodi e metodologie, sono in atto, seguiti e documentati.*

I risultati delle misurazioni saranno utilizzati nella funzione MANAGE *per assistere il monitoraggio del rischio e gli sforzi di risposta.*

Le pratiche relative alla misurazione dei rischi dell'IA saranno descritte nel Playbook NIST AI RMF.

La tabella 4 elenca le categorie e le sottocategorie della funzione MEASURE.

TABLE 4: CATEGORIES AND SUBCATEGORIES FOR THE MEASURE FUNCTION

CATEGORY	SUBCATEGORY
Funzione **MEASURE**: Identified risks are assessed, analyzed, or tracked	
MEASURE 1: Appropriate methods and metrics are identified and applied	*MEASURE 1.1: Approaches and metrics for quantitative or qualitative measurement of the most significant risks, identified by the outcome of the Map function, including context-relevant measures of trustworthiness are identified and selected for implementation. The risks or trustworthiness characteristics that will not be measured are properly documented.*
	MEASURE 1.2: Appropriateness of metrics and effectiveness of existing controls is regularly assessed and updated.
	MEASURE 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, and external stakeholders and affected communities are consulted in support of assessments.
MEASURE 2: Systems are evaluated for trustworthy characteristics.	*MEASURE 2.1: Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.*
	MEASURE 2.2: Evaluations involving human subjects comply with human subject protection requirements; and human subjects or datasets are representative of the intended population.
	MEASURE 2.3: System performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.
	MEASURE 2.4: Deployed product is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.
	MEASURE 2.5: AI system is evaluated regularly for safety. Deployed product is demonstrated to be safe and can fail safely and gracefully if it is made to operate beyond its knowledge limits. Safety metrics implicate system reliability and robustness, real-time monitoring, and response times for AI system failures.
	MEASURE 2.6: Computational bias is evaluated regularly and results are documented.
	MEASURE 2.7: AI system resilience and security is evaluated regularly and documented.
	MEASURE 2.8: AI model is explained, validated, and documented. AI system output is interpreted within its context and to inform responsible use and governance.
	MEASURE 2.9: Privacy risk of the AI system is examined regularly and documented.
	MEASURE 2.10: Environmental impact and sustainability of model training and management activities are assessed and documented.
MEASURE 3: Mechanisms for tracking identified risks over time are in place	*MEASURE 3.1: Approaches, personnel, and documentation are in place to regularly identify and track existing and emergent risks based on factors such as intended and actual performance in deployed contexts.*
	MEASURE 3.2: Risk tracking approaches are considered for settings where risks are difficult to assess using currently available measurement techniques or are not yet available.
MEASURE 4: Feedback about efficacy of measurement is gathered and assessed.	*MEASURE 4.1: Measurement approaches for identifying risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.*
	MEASURE 4.2: Measurement results regarding system trustworthiness in deployment context(s) are informed by domain expert and other stakeholder feedback to validate whether the system is performing consistently as intended. Results are documented.
	MEASURE 4.3: Measurable performance improvements (e.g., participatory methods) based on stakeholder consultations are identified and documented.

MANAGE

La funzione MANAGE *comporta l'allocazione delle risorse per la gestione del rischio ai rischi mappati e misurati su base regolare e come definito dalla funzione* GOVERN.

MANAGE *allinea gli aspetti tecnici della gestione del rischio ai suoi criteri e alle sue operazioni.*

Completata la funzione MANAGE, *saranno messi in atto piani per la prioritizzazione del rischio, per il monitoraggio e per il miglioramento continui.*

Le pratiche relative alla gestione dei rischi dell'IA saranno descritte nel Playbook NIST AI RMF.

Nella tabella 5 sono elencate le categorie e le sottocategorie della funzione Gestisci.

TABLE 5: CATEGORIES AND SUBCATEGORIES FOR THE MANAGE FUNCTION

CATEGORIA	SOTTOCATEGORIA
Funzione **MANAGE**: i rischi sono prioritari e attuati in base a un impatto previsto	
MANAGE 1: *i rischi dell'IA basati su valutazioni d'impatto e altri risultati analitici delle funzioni Mappa e Misura sono prioritari, risolti e gestiti*	MANAGE *1.1: Viene stabilito se il sistema di IA raggiunge lo scopo previsto e gli obiettivi dichiarati e deve procedere nello sviluppo o nella distribuzione.*
	MANAGE *1.2: Il trattamento dei rischi documentati ha la priorità in base all'impatto, alla probabilità e ai metodi delle risorse disponibili.*
	MANAGE *1.3: Le risposte ai rischi più significativi, identificati dalla funzione Map, sono sviluppate, pianificate e documentate. Le opzioni di risposta al rischio possono includere la mitigazione, il trasferimento, la condivisione, l'evitamento o l'accettazione.*
MANAGE *2: Le strategie per massimizzare i benefici e ridurre al minimo gli impatti negativi sono pianificate, preparate, implementate e documentate e informate dal contributo delle parti interessate*	MANAGE *2.1: Vengono prese in considerazione le risorse necessarie per gestire i rischi, insieme a sistemi, approcci o metodi alternativi praticabili e alla relativa riduzione della gravità dell'impatto o della probabilità di ogni potenziale azione.*
	MANAGE *2.2: I meccanismi sono in atto e applicati per sostenere il valore dei sistemi di IA distribuiti.*
	MANAGE *2.3: I meccanismi sono in atto e applicati per sostituire, disimpegnare o disattivare i sistemi di IA che dimostrano prestazioni o risultati incoerenti con l'uso previsto.*
MANAGE *3: Rischi da vengono gestite entità di terze parti*	MANAGE *3.1: i rischi provenienti da risorse di terze parti vengono regolarmente monitorati e i controlli dei rischi vengono applicati e documentati.*
MANAGE *4: Risposte a i rischi identificati e misurati sono documentati e monitorati regolarmente*	MANAGE *4.1: vengono implementati piani di monitoraggio del sistema post-distribuzione, inclusi meccanismi per acquisire e valutare il feedback degli utenti e delle parti interessate, l'appello e l'override, la disattivazione, la risposta agli incidenti e la gestione delle modifiche.*
	MANAGE *4.2: Le attività di miglioramento continuo misurabili sono integrate negli aggiornamenti del sistema e includono un regolare coinvolgimento degli stakeholder.*

DETTAGLIO FUNZIONI / CATEGORIE / SOTTOCATEGORIE

	CATEGORIA		CONSIDERAZIONI

Funzione: GOVERN

I processi di governance sono la spina dorsale della gestione del rischio e si concentrano sui potenziali impatti delle tecnologie di IA.

I risultati della funzione di governo promuovono una cultura della gestione del rischio all'interno delle organizzazioni che progettano, sviluppano, implementano o acquisiscono sistemi di IA (non solo per l'IA ma anche per le tradizionali).

Le categorie in questa funzione interagiscono tra loro e con altre funzioni, ma non si basano necessariamente su azioni precedenti.

Categoria: GOVERN-1

Politiche, processi, procedure e pratiche in tutta l'organizzazione relative alla mappatura, alla misurazione ed alla gestione dei rischi dell'IA sono in atto, trasparenti e implementate in modo efficace.

Sottocategoria: GOVERN 1.1

I requisiti legali e normativi che coinvolgono l'IA sono compresi, gestiti e documentati.

Alcuni requisiti legali impongono la documentazione e una maggiore trasparenza del sistema di IA.

Questi requisiti sono complessi e potrebbero non essere applicabili in tutti i contesti.

Ad esempio, i processi di test del sistema di IA per la misurazione della distorsione, come il trattamento disparato, non sono applicati in modo uniforme nel contesto legale.

Sottocategoria: GOVERN 1.2

- ✓ *Le caratteristiche di un'IA affidabile sono integrate nelle politiche, nei processi e nelle procedure organizzative.*
- ✓ *Le politiche organizzative dovrebbero:*
- ✓ *Definire termini e concetti chiave relativi ai sistemi di IA e l'ambito del loro uso previsto.*
- ✓ *Affrontare l'uso di dati sensibili o comunque rischiosi.*
- ✓ *Standard di dettaglio per la progettazione sperimentale, la qualità dei dati e l'addestramento del modello.*
- ✓ *Delineare e documentare processi e standard di mappatura e misurazione del rischio.*
- ✓ *Processi di verifica e convalida del modello di dettaglio.*
- ✓ *Processi di revisione dei dettagli per le funzioni legali e di rischio.*
- ✓ *Stabilire la frequenza e il dettaglio dei processi di monitoraggio, auditing e revisione.*
- ✓ *Delineare i requisiti di gestione del cambiamento.*
- ✓ *Delineare i processi per il coinvolgimento degli stakeholder interni ed esterni.*
- ✓ *Stabilire politiche per gli informatori per facilitare la segnalazione di gravi problemi del sistema di IA.*
- ✓ *Dettagliare e testare i piani di risposta agli incidenti.*
- ✓ *Verificare che le politiche formali di gestione del rischio dell'IA siano allineate agli standard legali esistenti e alle migliori pratiche e norme del settore.*
- ✓ *Stabilire politiche di gestione del rischio dell'IA che si allineano ampiamente alle caratteristiche affidabili del sistema dell'IA.*
- ✓ *Verificare che le politiche formali di gestione del rischio dell'IA includano i sistemi di IA attualmente implementati e di terze parti.*

Sottocategoria: GOVERN 1.3

Il processo di gestione del rischio e i suoi risultati sono stabiliti attraverso meccanismi trasparenti e tutti i rischi significativi sono misurati.

- ✓ *Stabilire e rivedere regolarmente le politiche di documentazione che riguardano le informazioni relative a:*
  - *Informazioni di contatto dell'attore AI*
  - *Giustificazione commerciale*
  - *Ambito e utilizzo*
  - *Presupposti e limitazioni*
  - *Descrizione dei dati di allenamento*
  - *Metodologia algoritmica*
  - *Approcci alternativi valutati*
  - *Descrizione dei dati di output*
  - *Risultati dei test e della convalida*
  - *Dipendenze a valle e a monte*
  - *Piani per l'implementazione, il monitoraggio e la gestione delle modifiche*
  - *Piani di coinvolgimento degli stakeholder*

✓ *Verificare che le politiche di documentazione per i sistemi di IA siano standardizzate in tutta l'organizzazione e aggiornate.*

✓ *Stabilire politiche per un sistema di inventario della documentazione modello e verificarne regolarmente la completezza, l'usabilità e l'efficacia.*

✓ *Stabilire meccanismi per riesaminare regolarmente l'efficacia dei processi di gestione del rischio.*

✓ *Identificare gli attori dell'IA responsabili della valutazione dell'efficacia dei processi e degli approcci di gestione del rischio e della correzione del corso sulla base dei risultati.*

Sottocategoria: **GOVERN 1.4**

Sono pianificati il monitoraggio continuo e la revisione periodica del processo di gestione del rischio e dei suoi esiti, con ruoli e responsabilità organizzative chiaramente definiti.

✓ *Stabilire politiche e procedure per il monitoraggio delle prestazioni del sistema di IA e per affrontare i pregiudizi e i problemi di sicurezza durante tutto il ciclo di vita del sistema.*

✓ *Stabilire politiche per la risposta agli incidenti del sistema di IA o confermare che le politiche di risposta agli incidenti esistenti si rivolgano ai sistemi di IA.*

✓ *Stabilire politiche per definire le funzioni organizzative e il personale responsabile del monitoraggio del sistema di IA e delle attività di risposta agli incidenti.*

✓ *Stabilire meccanismi per consentire la condivisione del feedback degli individui o delle comunità colpiti sugli impatti negativi dei sistemi di IA.*

✓ *Stabilire meccanismi per fornire ricorso agli individui o alle comunità colpiti per contestare i risultati problematici del sistema di IA.*

Categoria: **GOVERN-2**

Le strutture di responsabilità sono predisposte in modo che i team e gli individui appropriati siano autorizzati, responsabili e formati per mappare, misurare e gestire i rischi dell'IA.

Sottocategoria: **GOVERN 2.1**

I ruoli, le responsabilità e le linee di comunicazione relative alla mappatura, alla misurazione e alla gestione dei rischi dell'IA sono documentati e chiari a individui e team in tutta l'organizzazione.

✓ *Stabilire politiche che definiscano i ruoli e le responsabilità di gestione del rischio dell'IA per le posizioni direttamente e indirettamente correlate ai sistemi di IA, inclusi, a titolo esemplificativo:*

- *Consigli di amministrazione o comitati consultivi*
- *Alti dirigenti*
- *Funzioni di controllo dell'IA*
- *Gestione del prodotto*
- *Gestione del progetto*
- *Progettazione dell'IA*

- *Sviluppo dell'IA*
- *Interazione uomo-IA*
- *Test e valutazione dell'IA*
- *Acquisizione e approvvigionamento dell'IA*
- *Funzioni di valutazione dell'impatto*
- *Funzioni di supervisione*

✓ *Stabilire politiche che promuovano una comunicazione regolare tra gli attori dell'IA che partecipano agli sforzi di gestione del rischio dell'IA.*

✓ *Stabilire politiche che separino la gestione delle funzioni di sviluppo del sistema di IA dalle funzioni di test del sistema di IA, per consentire la correzione del corso indipendente dei sistemi di IA.*

✓ *Stabilire politiche per identificare, aumentare la trasparenza e prevenire i conflitti di interesse nella gestione del rischio dell'IA e per contrastare i bias di conferma e gli incentivi di mercato che possono ostacolare gli sforzi di gestione del rischio dell'IA.*

Sottocategoria: **GOVERN 2.2**

Al personale e ai partner dell'organizzazione viene fornita una formazione sulla gestione del rischio basata sull'IA per consentire loro di svolgere i propri compiti e responsabilità coerenti con le politiche, le procedure e gli accordi correlati.

✓ *Stabilire politiche per il personale che si occupa della formazione continua su:*

- *Leggi e regolamenti applicabili ai sistemi di IA.*
- *Impatti negativi che possono derivare dai sistemi di IA.*
- *Politiche dell'IA organizzative.*
- *Caratteristiche affidabili dell'IA.*

✓ *Verificare che le politiche organizzative dell'IA includano meccanismi che consentano al personale interno di riconoscere e impegnarsi nei propri ruoli e responsabilità.*

✓ *Verificare che le politiche organizzative rispondano alla gestione del cambiamento e includano meccanismi per comunicare e riconoscere le modifiche sostanziali del sistema di IA.*

✓ *Definire percorsi lungo le catene di responsabilità interne ed esterne per aumentare le apprensioni sui rischi.*

Sottocategoria: **GOVERN 2.3**

La leadership esecutiva dell'organizzazione considera le decisioni sui rischi associati allo sviluppo e all'implementazione del sistema di IA come loro responsabilità.

✓ *La gestione organizzativa può:*

- *Dichiarare la tolleranza al rischio per lo sviluppo o l'utilizzo di sistemi di IA.*
- *Supportare gli sforzi di gestione del rischio dell'IA e svolgere un ruolo attivo in tali sforzi.*
- *Supportare dirigenti competenti in materia di gestione del rischio.*
- *Delegare il potere, le risorse e l'autorizzazione per eseguire la gestione del rischio a ogni livello appropriato lungo la catena di gestione.*

✓ *Le organizzazioni possono istituire comitati consiliari per la gestione del rischio dell'IA e le funzioni di supervisione e integrare tali funzioni all'interno dei più ampi approcci di gestione del rischio aziendale dell'organizzazione.*

Categoria: **GOVERN-3**

I processi di diversità, equità, inclusione e accessibilità della forza lavoro hanno la priorità nella mappatura, misurazione e gestione dei rischi dell'IA durante tutto il ciclo di vita.

Sottocategoria: **GOVERN 3.1**

Il processo decisionale relativo alla mappatura, misurazione e gestione dei rischi dell'IA durante l'intero ciclo di vita è informato da un team diversificato dal punto di vista demografico e disciplinare, che include personale interno ed esterno. In particolare, i team che sono direttamente coinvolti nell'identificazione di considerazioni e rischi di progettazione includono una varietà di esperienze, competenze e background per garantire che i sistemi di IA soddisfino i requisiti al di là di un ristretto sottoinsieme di utenti.

✓ *Definire fin dall'inizio politiche e pratiche di assunzione che promuovano ruoli, competenze, abilità e capacità interdisciplinari per gli sforzi dell'IA.*

✓ *Definire politiche e pratiche di assunzione che portino alla diversità demografica e delle competenze di dominio; fornire al personale le risorse e il supporto necessari e facilitare il contributo del feedback e delle apprensioni del personale senza timore di rappresaglie.*

✓ *Stabilire politiche che facilitino l'inclusività e l'integrazione di nuove conoscenze nella pratica esistente.*

✓ *Ricercare competenze esterne per integrare la diversità organizzativa, l'equità, l'inclusione e l'accessibilità laddove manchino competenze interne.*

Categoria: **GOVERN-4**

I team organizzativi sono impegnati in una cultura che considera e comunica il rischio.

Sottocategoria: **GOVERN 4.1**

Sono in atto pratiche organizzative per promuovere un pensiero critico e una mentalità orientata alla sicurezza nella progettazione, sviluppo e implementazione di sistemi di IA per ridurre al minimo gli impatti negativi.

✓ *Stabilire politiche che richiedano l'inclusione di funzioni di supervisione (legale, conformità, gestione del rischio) fin dall'inizio del processo di progettazione del sistema.*

✓ *Stabilire politiche che promuovano un'efficace sfida alle decisioni di progettazione, implementazione e implementazione del sistema di IA, tramite meccanismi come le tre linee di difesa, gli audit dei modelli o il red-teaming, per garantire che i rischi sul posto di lavoro come il* PENSIERO DI GRUPPO *non prendano piede.*

✓ *Stabilire politiche che incentivino la mentalità orientata alla sicurezza e il pensiero critico generale e la revisione a livello organizzativo e procedurale.*

✓ *Stabilire protezioni per gli informatori per gli addetti ai lavori che segnalano problemi seri percepiti con i sistemi di IA.*

Sottocategoria: **GOVERN 4.2**

I team documentano i rischi e gli impatti della tecnologia che progettano, sviluppano o implementano e comunicano su questi impatti in modo più ampio.

✓ *Stabilire politiche e processi di VdI per i sistemi di IA utilizzati dall'organizzazione.*

✓ *Verificare che le politiche di VdI siano appropriate per valutare il potenziale impatto negativo di un sistema e la rapidità con cui un sistema cambia e che le valutazioni siano applicate regolarmente.*

✓ *Utilizzare le VdI per fornire valutazioni più ampie del rischio del sistema di IA.*

Sottocategoria: **GOVERN 4.3**

Sono in atto pratiche organizzative per consentire test, identificazione di incidenti e condivisione delle informazioni.

✓ *Stabilire politiche e procedure per facilitare ed equipaggiare i test del sistema di IA.*

✓ *Stabilire un impegno organizzativo per identificare i limiti del sistema di IA e condividere informazioni sui limiti all'interno di appropriati gruppi di attori di IA.*

✓ *Stabilire politiche per la risposta agli incidenti.*

✓ *Stabilire linee guida per la gestione e il controllo degli accessi relativi ai rischi e alle prestazioni del sistema di IA.*

Categoria: **GOVERN-5**

Sono in atto processi per un solido coinvolgimento degli stakeholder.

Sottocategoria: **GOVERN 5.1**

Sono in atto politiche e pratiche organizzative per raccogliere, considerare, assegnare priorità e integrare il feedback degli stakeholder esterni in merito ai potenziali impatti individuali e sociali relativi ai rischi dell'IA.

✓ *Stabilire politiche di gestione del rischio dell'IA che affrontino esplicitamente i meccanismi per raccogliere, valutare e incorporare il feedback degli stakeholder e degli utenti che potrebbero includere:*

- *Meccanismi di ricorso per uscite di sistema IA difettose.*
- *Ricompense di bug.*
- *Design incentrato sull'uomo.*
- *Interazione con l'utente e ricerca sull'esperienza.*
- *Coinvolgimento partecipativo delle parti interessate con individui e comunità che potrebbero subire impatti negativi.*

✓ *Verificare che il feedback delle parti interessate sia preso in considerazione e affrontato, comprese le apprensioni ambientali, e attraverso l'intera popolazione degli utenti previsti, comprese le popolazioni storicamente escluse, le persone con disabilità, gli anziani e quelle con accesso limitato a Internet e ad altre tecnologie di base.*

✓ *Chiarire i principi dell'organizzazione applicati ai sistemi di IA, considerando quelli che sono stati proposti pubblicamente, per informare gli stakeholder esterni dei valori dell'organizzazione. Prendi in considerazione la pubblicazione o l'adozione dei principi dell'IA.*

Sottocategoria: **GOVERN 5.2**

Vengono stabiliti meccanismi per consentire agli attori dell'IA d'incorporare regolarmente il feedback delle parti interessate aggiudicate nella progettazione e implementazione del sistema.

✓ *Riconoscere esplicitamente che i sistemi di IA e l'uso dell'IA presentano costi e rischi intrinseci insieme a potenziali benefici.*

✓ *Definire tolleranze di rischio ragionevoli per i sistemi di IA informati da leggi, regolamenti, migliori pratiche o standard di settore.*

✓ *Stabilire politiche che definiscano come assegnare i sistemi di IA ai livelli di tolleranza al rischio stabiliti combinando le valutazioni dell'impatto del sistema con la probabilità che si verifichi un impatto. Tale valutazione spesso comporta una combinazione di:*

- *Valutazioni econometriche degli impatti e delle probabilità di impatto per valutare il rischio del sistema di IA.*
- *Scale Rosso-Ambra-Verde (Red-Amber-Green [RAG]) per la gravità dell'impatto e la probabilità di valutare il rischio del sistema di IA.*
- *Definizione di politiche per l'allocazione delle risorse di gestione del rischio lungo livelli di tolleranza al rischio stabiliti, con sistemi a rischio più elevato che ricevono più risorse di gestione del rischio e supervisione.*
- *Definizione di politiche per l'approvazione, l'approvazione condizionale e la disapprovazione della progettazione, implementazione e implementazione dei sistemi di IA.*

✓ *Stabilire politiche che facilitino la disattivazione anticipata di un sistema di IA ritenuto rischioso al di là di una mitigazione pratica.*

Categoria: **GOVERN 6**

Sono in atto politiche e procedure chiare per affrontare i rischi dell'IA derivanti da software e dati di terze parti e altri problemi della catena di approvvigionamento.

Sottocategoria: **GOVERN 6.1**

Sono in atto le politiche e le procedure che affrontano i rischi associati a entità di terze parti.

✓ *Stabilire in modo collaborativo politiche che si rivolgono a sistemi e dati di IA di terze parti.*

✓ *Stabilire politiche relative a:*

- *Trasparenza nelle funzioni di sistema di terze parti, inclusa la conoscenza dei dati di addestramento, degli algoritmi di addestramento e di inferenza, nonché ipotesi e limitazioni.*
- *Test approfonditi di sistemi di IA di terze parti.*
- *Requisiti per istruzioni chiare e complete per l'utilizzo di sistemi di terze parti.*

Sottocategoria: **GOVERN 6.2**

Sono in atto processi di emergenza per gestire guasti o incidenti nei dati di terze parti o nei sistemi di IA ritenuti ad alto rischio.

- ✓ *Stabilire le politiche per la gestione degli errori di sistema di terze parti che includano la considerazione dei meccanismi di ridondanza per i sistemi di IA vitali di terze parti.*
- ✓ *Verificare che i piani di risposta agli incidenti rispondano a sistemi di IA di terze parti.*

Funzione: MAP

La funzione Map stabilisce il contesto e inquadra i rischi relativi a un sistema di IA.

Le informazioni raccolte in questa funzione informano le decisioni sulla gestione del modello, inclusa una decisione iniziale sull'adeguatezza o sulla necessità di una soluzione di IA.

Categoria: MAP 1

Sottocategoria: MAP 1.1

Sono compresi e documentati:

- ✓ *lo scopo previsto,*
- ✓ *le impostazioni potenziali in cui sarà implementato il sistema di IA,*
- ✓ *l'insieme o i tipi specifici di utenti insieme alle loro aspettative e gli impatti dell'uso del sistema.*

I presupposti e le relative limitazioni sullo scopo e sull'uso del sistema di IA sono enumerati, documentati e legati alle considerazioni sul TEVV e alle metriche di sistema.

- ✓ *Perseguire la progettazione del sistema di IA in modo mirato, dopo aver considerato le soluzioni non basate sull'IA.*
- ✓ *Definire e documentare l'attività, lo scopo, la funzionalità minima e i vantaggi del sistema di IA per valutare se vale la pena perseguire il progetto.*
- ✓ *Mantenere la consapevolezza degli standard di settore, tecnici e legali applicabili.*
- ✓ *Considerare in modo collaborativo le attività di progettazione del sistema di IA previste insieme a scopi imprevisti.*
- ✓ *Determinare i requisiti dell'utente e dell'organizzazione, inclusi i requisiti aziendali e tecnici.*
- ✓ *Determinare e delineare il contesto di utilizzo previsto e accettabile del sistema di IA, tra cui:*
  - ➢ *l'ambiente operativo,*
  - ➢ *gli impatti su individui, gruppi, comunità, organizzazioni e società,*
  - ➢ *le caratteristiche e compiti dell'utente,*
  - ➢ *l'ambiente sociale.*
- ✓ *Tenere traccia e documentare i sistemi di IA esistenti detenuti dall'organizzazione e quelli gestiti o supportati da entità di terze parti.*
- ✓ *Acquisire e mantenere la consapevolezza sulla valutazione delle affermazioni scientifiche relative alle prestazioni e ai vantaggi del sistema di IA prima di avviare la progettazione del sistema.*
- ✓ *Identificare l'interazione e/o i ruoli uomo-IA, ad esempio se l'applicazione supporterà o sostituirà il processo decisionale umano.*
- ✓ *Pianificare i rischi relativi alle configurazioni dell'IA umana e documentare i requisiti, i ruoli e le responsabilità per la supervisione umana dei sistemi implementati.*

Sottocategoria: MAP 1.2

Gli attori interdisciplinari dell'IA, le competenze, le abilità e le capacità di stabilire un contesto riflettono la diversità demografica e l'ampio dominio e l'esperienza dell'utente, e la loro partecipazione è documentata.

Le opportunità di collaborazione interdisciplinare sono prioritarie.

- ✓ *Istituire team interdisciplinari per riflettere un'ampia gamma di abilità, competenze e capacità per gli sforzi dell'IA.*

Verifica che l'appartenenza al team includa sia la diversità demografica, l'ampia esperienza nel settore che le esperienze vissute.

Composizione del team di documenti.

✓ *Creare e consentire a gruppi di esperti interdisciplinari di acquisire, apprendere e coinvolgere le interdipendenze dei sistemi di IA implementati e delle relative terminologie e concetti da discipline al di fuori della pratica dell'IA come diritto, sociologia, psicologia, antropologia, politiche pubbliche, progettazione di sistemi e ingegneria.*

Sottocategoria: **MAP 1.3**

Il valore aziendale o il contesto di utilizzo aziendale è stato chiaramente definito o, nel caso di valutazione dei sistemi di IA esistenti, rivalutato.

✓ *Creare pratiche trasparenti nei processi di sviluppo del sistema di IA.*

✓ *Rivedere lo scopo del sistema documentato da una prospettiva socio-tecnica e in considerazione dei valori sociali.*

✓ *Determinare il possibile disallineamento tra i valori sociali e i principi organizzativi dichiarati e il codice etico.*

✓ *Segnalare incentivi latenti che possono contribuire a impatti negativi.*

✓ *Bilanciare lo scopo del sistema di IA con potenziali rischi, valori sociali e principi organizzativi dichiarati.*

Sottocategoria: **MAP 1.4**

La missione dell'organizzazione e gli obiettivi rilevanti per la tecnologia AI sono stati compresi.

✓ *Riconciliare le apprensioni documentate sul contesto di utilizzo o lo scopo del sistema con i valori dichiarati dell'organizzazione, le dichiarazioni di intenti, gli impegni di responsabilità sociale e i principi dell'IA.*

✓ *Riconsiderare la progettazione, la strategia di implementazione o l'implementazione di sistemi di IA con potenziali impatti che non riflettono i valori istituzionali.*

Sottocategoria: **MAP 1.5**

Sono determinate le tolleranze al rischio organizzativo.

✓ *Stabilire livelli di tolleranza al rischio per i sistemi di IA e allocare le risorse di supervisione appropriate a ciascun livello.*

✓ *Identificare le soglie di rischio massime consentite al di sopra delle quali il sistema non verrà implementato, o dovrà essere disattivato prematuramente, all'interno dell'impostazione contestuale o dell'applicazione.*

✓ *I tentativi di utilizzare un sistema per scopi "off-label" dovrebbero essere affrontati con cautela, specialmente in contesti che le organizzazioni hanno ritenuto ad alto rischio.*

✓ *Documentare le decisioni, i compromessi relativi al rischio e le limitazioni del sistema.*

Sottocategoria: **MAP 1.6**

Le pratiche e il personale per le attività di progettazione consentono un impegno regolare con le parti interessate e integrano il feedback fruibile degli utenti e della comunità sugli impatti negativi imprevisti.

✓ *Mantenere la consapevolezza e la documentazione degli individui, dei gruppi o delle comunità che compongono gli stakeholder interni ed esterni del sistema.*

✓ *Verificare che le competenze e le pratiche appropriate siano disponibili internamente per svolgere attività di coinvolgimento degli stakeholder come raccogliere, acquisire e sintetizzare il feedback degli stakeholder e tradurlo per le funzioni di progettazione e sviluppo dell'IA.*

✓ *Stabilire meccanismi per la comunicazione e il feedback regolari tra i pertinenti attori dell'IA e le parti interessate interne o esterne in relazione alla progettazione del sistema o alle decisioni di implementazione.*

✓ *Definire quali attori dell'IA, oltre ai team di progettazione e sviluppo dell'IA, esamineranno le attività di progettazione, implementazione e funzionamento del sistema.*

✓ *Definisci quali attori dell'IA amministreranno e implementeranno attività di test, valutazione, verifica e convalida (TEVV) durante il ciclo di vita dell'IA.*

Sottocategoria: **MAP 1.7**

I requisiti di sistema (ad esempio, "il sistema deve rispettare la privacy dei suoi utenti") sono elicitati e compresi dalle parti interessate. Le decisioni di progettazione tengono conto delle implicazioni socio-tecniche per affrontare il rischio dell'IA.

- ✓ *Incorpora in modo proattivo caratteristiche affidabili nei requisiti di sistema.*
- ✓ *Considerare i fattori di rischio relativi alle configurazioni e alle attività Human-AI.*
- ✓ *Analizzare le dipendenze tra fattori contestuali e requisiti di sistema.*
- ✓ *Elencare gli impatti che possono derivare dal non considerare pienamente l'importanza delle caratteristiche di affidabilità in qualsiasi processo decisionale.*
- ✓ *Seguire le tecniche di progettazione responsabile in attività quali l'ingegneria del software, la gestione dei prodotti e l'impegno partecipativo.*
- ✓ *Alcuni esempi per sollecitare e documentare i requisiti delle parti interessate includono documenti sui requisiti del prodotto (PRD), storie degli utenti, ricerca sull'interazione dell'utente/esperienza dell'utente (UI/UX), ingegneria dei sistemi, etnografia e metodi sul campo correlati.*
- ✓ *Condurre ricerche sugli utenti per comprendere individui, gruppi e comunità che saranno influenzati dall'IA, i loro valori e contesto e il ruolo dei pregiudizi sistemici e storici.*
- ✓ *Integrare gli apprendimenti nelle decisioni sulla selezione e rappresentazione dei dati.*

Categoria: **MAP 2**

È eseguita la classificazione del sistema di IA.

Sottocategoria: **MAP 2.1**

È definita la fase specifica, con i relativi metodi usati, per implementare l'attività che il sistema di IA supporterà (ad es. classificatori, modelli generativi, raccomandatori, ecc.).

Definire e documentare le attività di apprendimento esistenti e potenziali del sistema di IA insieme a presupposti e limiti noti.

Sottocategoria: **MAP 2.2**

Le informazioni sono documentate sui limiti di conoscenza del sistema e su come l'output sarà utilizzato e supervisionato dagli esseri umani.

- ✓ *Estendere la documentazione oltre i requisiti di sistema e attività per includere i possibili rischi dovuti ai contesti di distribuzione e alle configurazioni dell'IA umana.*
- ✓ *Seguire i processi di feedback degli stakeholder per determinare se un sistema ha raggiunto il suo scopo documentato in un determinato contesto di utilizzo e se gli utenti possono comprendere correttamente gli output o i risultati del sistema.*
- ✓ *Documentare le dipendenze dai dati a monte e altri sistemi di IA, anche se il sistema specificato è una dipendenza a monte per un altro sistema di IA o altri dati.*
- ✓ *Documentare le connessioni che il sistema di IA oi dati dovranno avere con reti esterne (incluso Internet), mercati finanziari e infrastrutture critiche che possono presentare esternalità negative.*
- ✓ *Identificare e documentare gli impatti negativi nell'ambito della considerazione delle soglie di rischio più ampie e della successiva distribuzione go/no-go, nonché delle decisioni di disattivazione post-distribuzione.*

Sottocategoria: **MAP 2.3**

Le considerazioni sull'integrità scientifica e sul TEVV sono identificate e documentate, comprese quelle relative alla progettazione sperimentale, alla raccolta e alla selezione dei dati (ad esempio, disponibilità, rappresentatività, idoneità) e alla convalida del costrutto.

- ✓ *Documentare le ipotesi fatte e le tecniche utilizzate durante la selezione, la cura, la preparazione e l'analisi dei dati, e quando si identificano costrutti e obiettivi proxy e si sviluppano indici, specialmente quando si cerca di misurare concetti che sono intrinsecamente non osservabili (ad es. "noleggio", "criminalità", "prestazione").*
- ✓ *Mappare l'aderenza alle politiche che affrontano i dati e costruiscono validità, pregiudizi, privacy e sicurezza per i sistemi di IA e verificano la documentazione, la supervisione e i processi.*

✓ *Stabilire processi e pratiche che utilizzino tecniche di progettazione sperimentale per la raccolta, la selezione e le pratiche di gestione dei dati.*

✓ *Stabilire pratiche per garantire che i dati utilizzati nei sistemi di IA siano collegati allo scopo documentato del sistema di IA (ad esempio, mediante metodi di scoperta causale).*

✓ *Stabilire e documentare processi per garantire che la derivazione dei dati di test e formazione sia ben compresa, tracciabile e che le risorse di metadati siano disponibili per la mappatura dei rischi.*

✓ *Documentare i limiti noti, gli sforzi di mitigazione del rischio associati e i metodi utilizzati per la raccolta, la selezione, l'etichettatura, la pulizia e l'analisi dei dati di formazione (ad es. trattamento di dati mancanti, spuri o anomali; stimatori distorti).*

✓ *Stabilire e documentare le pratiche per verificare la presenza di capacità che superano quelle pianificate, come le proprietà emergenti, e per rivedere le precedenti fasi di gestione del rischio alla luce di eventuali nuove capacità.*

✓ *Stabilire processi per testare e verificare che le ipotesi di progettazione sull'insieme dei contesti di distribuzione continuino a essere accurate e sufficientemente complete.*

✓ *Collaborare con esperti di dominio per:*

- *Acquisire e mantenere la consapevolezza e la conoscenza contestuale su come il comportamento umano si riflette nei set di dati, nei fattori e nelle dinamiche organizzative e nella società.*
- *Identificare approcci partecipativi per configurazioni responsabili dell'IA e compiti di supervisione, tenendo conto delle fonti di bias cognitivo.*
- *Identificare le tecniche per gestire e mitigare le fonti di bias (sistemiche, computazionali, umano-cognitive) nei modelli e nei sistemi computazionali, e le ipotesi e le decisioni nel loro sviluppo.*

✓ *Seguire i principi statistici standard e documentare la misura in cui la tecnologia proposta non soddisfa i criteri di convalida standard.*

✓ *Indagare e documentare potenziali impatti negativi dovuti a problemi della catena di approvvigionamento che possono entrare in conflitto con valori e principi organizzativi.*

Categoria: **MAP 3**

Sono comprese le capacità dell'IA, l'obiettivo utilizzato, gli obiettivi, i benefici attesi e la costa rispetto allo status quo.

Sottocategoria: **MAP 3.1**

I vantaggi della funzionalità e delle prestazioni del sistema previste sono esaminati e documentati.

✓ *Utilizzare approcci partecipativi e interagire con gli utenti finali del sistema per valutare l'efficacia del sistema e l'interpretabilità dell'output delle attività di IA.*

✓ *Incorporare il feedback degli stakeholder sui benefici percepiti del sistema al di là dello status quo.*

✓ *Allineare i requisiti di sistema con lo scopo previsto e documentare le decisioni.*

✓ *Eseguire l'analisi del contesto relativa al periodo di tempo, ai problemi di sicurezza, all'area geografica, all'ambiente fisico, agli ecosistemi, all'ambiente sociale e alle norme culturali all'interno dell'ambiente previsto (o condizioni che si avvicinano molto all'ambiente previsto).*

Sottocategoria: **MAP 3.2**

I potenziali costi, inclusi i costi non monetari, che derivano da errori previsti o realizzati o dalle prestazioni del sistema sono esaminati e documentati.

✓ *Eseguire un'analisi del contesto per mappare gli impatti negativi derivanti dalla mancata integrazione delle caratteristiche di affidabilità.*

Quando gli impatti negativi non sono diretti o evidenti, gli attori dell'IA dovrebbero impegnarsi con le parti interessate esterne per indagare e documentare:

- *Chi potrebbe essere danneggiato?*
- *Cosa potrebbe essere danneggiato?*
- *Quando potrebbe verificarsi un danno?*
- *Come potrebbe sorgere un danno?*

✓ *Implementare procedure per valutare regolarmente i costi qualitativi e quantitativi dei guasti interni ed esterni del sistema di IA.*

✓ *Sviluppare azioni per prevenire, rilevare e/o correggere potenziali rischi e relativi impatti.*

✓ *Valutare regolarmente i costi dei guasti per prendere decisioni di distribuzione go/no-go durante tutto il ciclo di vita del sistema di IA.*

Sottocategoria: **MAP 3.3**

L'ambito dell'applicazione mirata è specificato, ristretto e documentato in base al contesto stabilito e alla classificazione del sistema di IA.

✓ *Considerare di restringere i contesti per l'implementazione del sistema, inclusi i fattori relativi a:*

- *In che modo i risultati possono avere un impatto diretto o indiretto sugli utenti e le parti interessate.*
- *Tempo di implementazione del sistema tra un riaddestramento e l'altro.*
- *Regioni geografiche in cui opera il sistema.*

✓ *Coinvolgi gli attori dell'IA dalle funzioni legali e di approvvigionamento quando specifichi l'ambito dell'applicazione target.*

Categoria: **MAP 4**

I rischi e i vantaggi sono mappati per software e dati di terze parti.

Sottocategoria: **MAP 4.1**

Gli approcci per la mappatura dei rischi tecnologici di terze parti sono in atto e documentati.

✓ *Esaminare i rapporti di audit, i risultati dei test, le percorsi dei prodotti, le garanzie, i termini di servizio, i contratti di licenza con l'utente finale, i contratti e altra documentazione relativa a entità di terze parti per assistere nella valutazione del valore e nelle attività di gestione del rischio.*

✓ *Esaminare i programmi di rilascio del software di terze parti e i piani di gestione delle modifiche del software (hotfix, patch, aggiornamenti, garanzie di compatibilità futura e precedente) per individuare eventuali irregolarità che possono contribuire ai rischi del sistema di IA.*

✓ *Inventario del materiale di terze parti (hardware, software open source, modelli di base, dati open source, software proprietario, dati proprietari, ecc.) necessario per l'implementazione e la manutenzione del sistema.*

✓ *Rivedere i licenziamenti relativi alla tecnologia e al personale di terze parti per valutare i potenziali rischi dovuti alla mancanza di un supporto adeguato.*

Sottocategoria: **MAP 4.2**

I controlli dei rischi interni relativo ai rischi tecnologici di terze parti sono in atto e documentati.

✓ *Fornire risorse come modelli di documentazione del modello ed elenchi di software attendibili per assistere nell'inventario tecnologico di terze parti e nelle attività di approvazione.*

✓ *Esaminare il materiale di terze parti (inclusi dati e modelli) per i rischi relativi a pregiudizi, privacy dei dati e vulnerabilità della sicurezza.*

✓ *Applicare i controlli, come l'approvvigionamento, la sicurezza e i controlli sulla privacy dei dati, a tutte le tecnologie di terze parti acquisite.*

Categoria: **MAP 5**

È possibile accedere agli impatti su individui, gruppi, comunità, organizzazioni o società.

Sottocategoria: **MAP 5.1**

I potenziali impatti positivi o negativi su individui, gruppi, comunità, organizzazioni o società sono regolarmente identificati e documentati.

✓ *Stabilire e documentare i processi di coinvolgimento degli stakeholder nelle prime fasi della formulazione del sistema per identificare i potenziali impatti del sistema su individui, gruppi, comunità, organizzazioni e società.*

✓ *Impiegare metodi come la* PROGETTAZIONE SENSIBILE AL VALORE (VALUE SENSITIVE DESIGN - VSD) *per identificare i disallineamenti tra i valori organizzativi e sociali e l'implementazione e l'impatto del sistema.*

✓ *Identificare approcci per coinvolgere, acquisire e incorporare input dagli utenti del sistema e da altri stakeholder chiave per assistere con il monitoraggio continuo degli impatti e dei rischi emergenti.*

✓ *Incorporare metodi quantitativi, qualitativi e misti nella valutazione e documentazione dei potenziali impatti su individui, gruppi, comunità, organizzazioni e società.*

✓ *Identificare un team (interno o esterno) indipendente dalle funzioni di progettazione e sviluppo dell'IA per valutare i vantaggi del sistema, gli impatti positivi e negativi e la loro probabilità.*

✓ *Sviluppare procedure di VdI che incorporino elementi e metodi socio-tecnici e pianificare la normalizzazione attraverso la cultura organizzativa.*

✓ *Riesaminare e perfezionare regolarmente i processi di VdI.*

Sottocategoria: **MAP 5.2**

Sono identificate e documentate la probabilità e l'entità di ciascun impatto identificato in base all'uso previsto, agli usi passati dei sistemi di IA in contesti simili, alle segnalazioni di incidenti pubblici, al feedback degli stakeholder o ad altri dati.

✓ *Stabilire scale di valutazione per misurare l'impatto del sistema di IA. Le scale possono essere qualitative, come il Rosso-Ambra-Verde (Red-Amber-Green [RAG]), o possono comportare simulazioni o approcci econometrici. Documentare e applicare le scale in modo uniforme nel portafoglio di IA dell'organizzazione.*

✓ *Applicare regolarmente valutazioni d'impatto nelle fasi chiave del ciclo di vita dell'IA, connesse agli impatti del sistema e alla frequenza degli aggiornamenti del sistema.*

✓ *Valutare i vantaggi e gli impatti negativi del sistema in relazione a caratteristiche affidabili.*

Sottocategoria: **MAP 5.3**

Le valutazioni dei benefici rispetto agli impatti si basano su analisi di impatto, di entità e di probabilità del rischio.

✓ *Rivedere ed esaminare la documentazione, inclusi lo scopo e i vantaggi del sistema, e mappare i potenziali impatti con le probabilità associate.*

✓ *Documentare il rischio stimato del sistema.*

✓ *Effettuare una determinazione "go/no go" in base all'entità e alla probabilità di impatto.*

✓ *Non implementare (no-go) o disattivare il sistema se il rischio stimato supera le tolleranze o le soglie dell'organizzazione.*

✓ *Se viene presa la decisione di procedere con l'implementazione, assegnare al sistema un'adeguata tolleranza al rischio e allineare le risorse di supervisione al rischio valutato.*

5 – I QUATTRO PRINCIPI D'ESPLICAZIONE DELL'IA [NIST IR 8312]

INTRODUZIONE

I sistemi di IA dovrebbero fornire prove di accompagnamento a risultati e processi; fornire spiegazioni comprensibili ai singoli utenti; fornire spiegazioni che riflettano correttamente il processo del sistema per generare il risultato; e che un sistema funziona solo nelle condizioni per le quali è stato progettato e quando raggiunge una sufficiente fiducia nei suoi risultati.

I QUATTRO *per l'*ESPLICAZIONE *dell'IA comprendenti le proprietà fondamentali per i sistemi sono:*

1. EXPLANATION,
2. MEANINGFUL,
3. EXPLANATION ACCURACY,
4. KNOWLEDGE LIMITS.

*Attraverso un significativo coinvolgimento delle parti interessate, questi quattro principi sono stati sviluppati per comprendere (*COMPRENSIBILITÀ*) la natura multidisciplinare dell'IA spiegabile, compresi i campi dell'informatica, dell'ingegneria e della psicologia.*

In termini di accettazione e fiducia della società, gli sviluppatori di sistemi IA potrebbero dover considerare che più attributi di un sistema IA possano influenzare la percezione pubblica del sistema.

La COMPRENSIBILITÀ *dell'IA è una delle numerose proprietà che caratterizzano la fiducia nei sistemi IA. Altre proprietà includono* RESILIENZA, AFFIDABILITÀ, DISTORSIONE *e* RESPONSABILITÀ.

Dal punto di vista informatico, collochiamo gli algoritmi e i sistemi IA spiegabili esistenti nel contesto di questi quattro principi.

Da un punto di vista psicologico, indaghiamo su quanto bene le spiegazioni delle persone seguano i nostri quattro principi.

5.1 – DESCRIZIONE DEI QUATTRO PRINCIPI D'ESPLICAZIONE DELL'IA

Questi principi sono fortemente influenzati dalla considerazione dell'interazione del sistema IA con il destinatario umano delle informazioni.

I requisiti della situazione data, il compito a portata di mano e il consumatore influenzeranno il tipo di spiegazione ritenuta appropriata per la situazione. Questi possono includere, ma non sono limitati a, requisiti normativi e legali, controllo di qualità di un sistema IA e relazioni con i clienti.

I nostri quattro principi hanno lo scopo di catturare un ampio insieme di motivazioni, ragioni e prospettive.

L'output di un sistema IA è il risultato di una query al medesimo sistema e varia in base alle attività svolte.

- ✓ *Per una richiesta di prestito è un esempio in cui l'output è una decisione: approvata o negata.*
- ✓ *Per un sistema di raccomandazione, l'output potrebbe essere un elenco di film consigliati.*
- ✓ *Per un sistema di controllo grammaticale, l'output è errori grammaticali e correzioni consigliate.*

In breve, i nostri quattro principi di ESPLICAZIONE *dell'IA sono:*

1. SPIEGAZIONE [EXPLANATION]: *i sistemi forniscono prove o motivazioni di accompagnamento per tutti i risultati ottenuti.*
2. SIGNIFICATIVO [MEANINGFUL]: *i sistemi forniscono spiegazioni comprensibili ai singoli utenti.*
3. ACCURATEZZA SPIEGAZIONE [EXPLANATION ACCURACY]: *la spiegazione riflette correttamente il processo del sistema per la cancellazione dell'output.*
4. LIMITI DI CONOSCENZA [KNOWLEDGE LIMITS]: *il sistema funziona solo in condizioni per le quali è stato progettato o quando il sistema raggiunge una sufficiente fiducia nella sua produzione.*

Questi sono definiti e contestualizzati in modo più dettagliato di seguito.

La figura 1 mostra i principi e indica che, per essere considerato spiegabile, un sistema deve prima avere una spiegazione o contenere prove di accompagnamento a cui è possibile accedere.

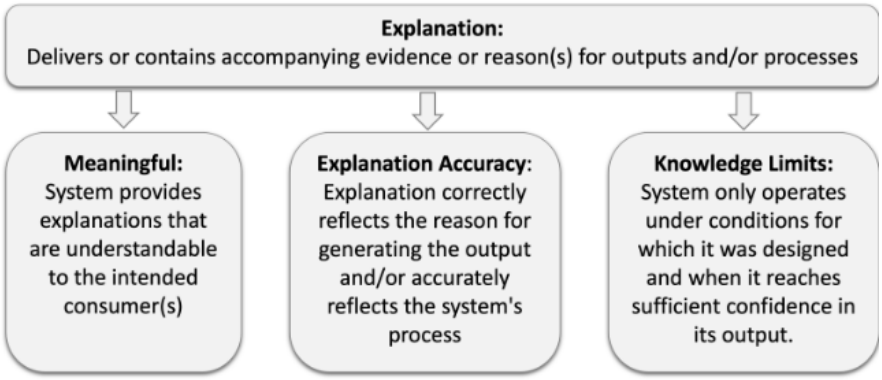


FIG. 1. I QUATTRO PRINCIPI DELL'ESPLICABILITÀ DELL'IA

Le frecce indicano che, per essere esplicabile, un sistema deve fornire una comprensione e una giustificazione.

I restanti **tre principi** sono le **proprietà** fondamentali di tali **spiegazioni**.

SPIEGAZIONE [EXPLANATION]

Il principio di SPIEGAZIONE *obbliga i sistemi a fornire prove, supporto o ragionamento per ogni risultato.*

Di per sé, questo principio non richiede che le prove siano corrette, informative o intelligibili; si limita ad affermare che un sistema è in grado di fornire una spiegazione.

Questo principio non impone alcuna metrica di qualità a tali spiegazioni.

SIGNIFICATIVO [MEANINGFUL]

Un sistema soddisfa il principio SIGNIFICATIVO *se il destinatario previsto comprende le spiegazioni del sistema.*

Ci sono punti in comune tra le spiegazioni che possono renderle più significative. Ad esempio, affermare perché il sistema si è comportato in un certo modo può essere più comprensibile anziché descrivere il "perché" non si è comportato in un certo modo.

Diversi fattori influenzano le informazioni che le persone troveranno importanti, rilevanti o utili. Questi includono le conoscenze e le esperienze precedenti di una persona e le differenze psicologiche complessive tra le persone. Inoltre, ciò che considerano significativo cambierà nel tempo man mano che acquisiscono esperienza con un'attività o un sistema.

Il principio SIGNIFICATIVO *sarà raggiunto comprendendo i bisogni del pubblico, il livello di competenza e la pertinenza alla domanda o alla domanda in questione.*

La sfida è sviluppare protocolli di misurazione che si adattino a diversi tipi di pubblico.

Piuttosto che considerarlo un peso, sosteniamo che sia la consapevolezza sia l'apprezzamento del contesto di una spiegazione supportino la capacità di misurare la qualità delle spiegazioni dell'IA.

ACCURATEZZA DELLA SPIEGAZIONE [EXPLANATION ACCURACY]

I principi SPIEGAZIONE *e* SIGNIFICATO *richiedono un metodo per produrre spiegazioni intelligibili al pubblico.*

Questi due principi non richiedono che una spiegazione rifletta correttamente il processo di un sistema per generare il suo output.

*Il principio dell'*ACCURATEZZA *della spiegazione impone la veridicità alle spiegazioni di un sistema.*

*L'*ACCURATEZZA *della* SPIEGAZIONE *è un concetto distinto dall'*ACCURATEZZA *della* DECISIONE.

L'ACCURATEZZA della DECISIONE si riferisce al fatto che il giudizio del sistema sia corretto o errato.

Indipendentemente dall'accuratezza della decisione del sistema, la spiegazione corrispondente può o non può descrivere accuratamente come il sistema è giunto alla sua conclusione o azione.

*L'*ACCURATEZZA *della* SPIEGAZIONE *deve tenere conto del livello di dettaglio nella spiegazione.*

Questo è simile al modo in cui gli esseri umani si avvicinano alla spiegazione di argomenti complessi. Un professore di neuroscienze può spiegare a un collega una nuova scoperta con dettagli estesi e tecnici. Lo stesso risultato sarà probabilmente distillato e modificato per essere presentato a uno studente universitario al fine di presentare i dettagli pertinenti e di livello superiore. Quello stesso professore potrebbe spiegare la scoperta in modo molto diverso ai loro amici e genitori non addestrati.

*Insieme, questo evidenzia il punto in cui l'*ACCURATEZZA *e la* SIGNIFICATIVITÀ *della spiegazione interagiscono.*

5.2 - LIMITI DI CONOSCENZA [KNOWLEDGE LIMITS]

I principi precedenti presuppongono implicitamente che un sistema operi nell'ambito della sua progettazione e dei suoi LIMITI *di* CONOSCENZA.

Il principio dei LIMITI *di* CONOSCENZA *afferma che i sistemi identificano i casi in cui non sono stati progettati o approvati per operare, o nei casi per i quali le loro risposte non sono affidabili.*

Identificando e dichiarando i limiti di conoscenza, questa pratica salvaguarda le risposte in modo che non venga fornito un giudizio quando potrebbe essere inappropriato farlo.

Questo principio può aumentare la fiducia in un sistema prevenendo risultati fuorvianti, pericolosi o ingiusti.

Ci sono due modi in cui un sistema può raggiungere o superare i suoi limiti di conoscenza. In un certo senso, l'operazione o la query al sistema può essere al di fuori del suo dominio.

Ad esempio, in un sistema creato per classificare le specie di uccelli, un utente può inserire l'immagine di una mela. Il sistema potrebbe restituire una risposta per indicare che non è stato possibile trovare uccelli nell'immagine di input; pertanto, il sistema non può fornire una risposta.

Questa è sia risposta sia spiegazione.

Riprendendo l'esempio del sistema di classificazione degli uccelli, l'immagine di input di un uccello potrebbe essere troppo sfocata per determinarne la specie. In questo caso, il sistema potrebbe riconoscere che l'immagine è di un uccello ma che l'immagine è di bassa qualità. Un esempio di output potrebbe essere: "Ho trovato un uccello nell'immagine, ma la qualità dell'immagine è troppo bassa per identificarlo".

5.3 - FINALITÀ E STILI DELLE SPIEGAZIONI

Per illustrare l'ampia gamma di spiegazioni, caratterizziamo le spiegazioni in base a due proprietà:

1) *lo* SCOPO *e*
2) *lo* STILE.

1) *Lo* SCOPO *è il motivo per cui una persona richiede la spiegazione o a quale domanda la spiegazione intende rispondere.*
2) *Lo* STILE *descrive come viene fornita la spiegazione.*

Lo scopo della spiegazione influenzerà a sua volta il suo stile.

TRE ELEMENTI DI STILE

Nella Figura 2, visualizziamo i nostri tre elementi di stile:

1) livello di dettaglio,
2) grado di interazione tra l'uomo e la macchina,
3) il suo formato.

Questi elementi sono strettamente correlati al rispetto dei quattro principi.

Pertanto, considerare questi getterà le basi per la produzione di spiegazioni.

Di seguito li esponiamo più in dettaglio.

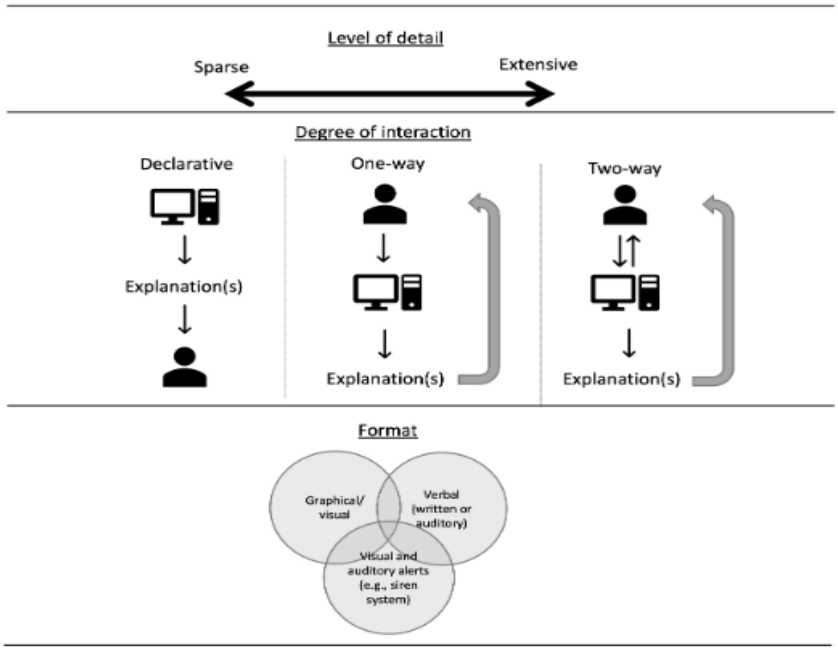


Fig. 2. Illustration of our elements of explanation styles.

Il livello di dettaglio è rappresentato come un intervallo, da **CARENTE** a **ESTESA**.

Per **CARENTE**, intendiamo che la quantità di informazioni fornite è breve, limitata e/o di alto livello, priva di dettagli.

Un esempio di spiegazione sparsa potrebbe essere una spiegazione per una decisione presa da un sistema di allerta (ad esempio, "processi di sistema rallentati a causa del surriscaldamento.").

Una spiegazione **ESTESA** può contenere informazioni dettagliate su un sistema e/o fornire una grande quantità di informazioni (ad esempio, un report con informazioni di sistema rilevanti per comprenderne il processo).

Mettiamo il grado di interazione UOMO-MACCHINA *in tre categorie:*

1) SPIEGAZIONI DICHIARATIVE,
2) INTERAZIONE UNIDIREZIONALE e
3) INTERAZIONE BIDIREZIONALE.

In una SPIEGAZIONE DICHIARATIVA, *i sistemi forniscono una spiegazione e non vi sono ulteriori interazioni.*

Questo descrive i metodi di IA spiegabili più attuali.

Ad esempio, un sistema di richiesta di prestito può sempre produrre la motivazione per un'accettazione o un rifiuto.

Un classificatore di oggetti può produrre una mappa di salienza.

Una scheda modello può contenere informazioni predeterminate sul sistema.

Una spiegazione dichiarativa si basa su una query predefinita, ad esempio "perché il classificatore di oggetti ha prodotto questa decisione?".

L'essere umano non può alterare la domanda posta (salvo un cambiamento nel sistema stesso per produrre qualcosa di diverso).

*Un grado più elevato di interazione è un'*INTERAZIONE UNIDIREZIONALE.

Per questo, la spiegazione è determinata sulla base di una query o di un input di domanda al sistema.

Ad esempio, questo potrebbe essere un output grafico a seconda dei fattori che una persona desidera visualizzare. Ciò può consentire al consumatore della spiegazione la possibilità di sondare ulteriormente o di inviare domande diverse.

*Definiamo la categoria con il livello di interazione più profondo come l'*INTERAZIONE BIDIREZIONALE.

Questo modella una conversazione tra le persone.

La persona può sondare ulteriormente e la macchina può sondare indietro, porre domande chiarificatrici o fornire nuove strade di esplorazione.

Ad esempio, un sistema può sondare l'utente per ulteriori dettagli o proporre domande alternative.

A nostra conoscenza, le interazioni bidirezionali non esistono ancora.

Il formato della spiegazione include avvisi visivi e grafici, verbali e uditivi o visivi.

Esempi di formati grafici includono output di analisi dei dati o mappe di salienza.

I formati verbali possono includere risultati e resoconti scritti, nonché output sonori, come il parlato.

Un'altra forma di spiegazione può catturare l'attenzione di un pubblico inconsapevole.

Una sirena o un sistema di luci può produrre diversi allarmi, schemi di lampeggio della luce e/o colori della luce come spiegazione che allertano il pubblico. Ad esempio, un tono o uno schema specifico della sirena potrebbe indicare qualcosa sullo stato di un sistema che potrebbe richiedere attenzione.

Ciascuno di questi elementi di stile dovrà essere considerato per produrre una spiegazione per il suo scopo e per soddisfare i quattro principi.

Alcuni casi possono richiedere una spiegazione semplice e dichiarativa come lo stile più appropriato per ottimizzare quanto sia significativo. Questo a volte è il caso di un'emergenza meteorologica, come quando un tornado si trova nell'area.

5.4 - GESTIONE DEL RISCHIO D'ESPLICAZIONE DELL'IA

Il rischio è definito come "L'EFFETTO DELL'INCERTEZZA DEGLI OBIETTIVI" *e comprende sia esiti negativi* (minacce) *sia risultati positivi* (opportunità).

La gestione del rischio è un processo che può essere utilizzato per definire, valutare e mitigare il rischio.

L'ESPLICAZIONE *può mitigare i rischi dell'IA valutando, misurando o prevedendo il rischio in un modello o sistema.*

Le spiegazioni possono essere utilizzate per testare le vulnerabilità.

Tuttavia, l'IA "spiegabile" può introdurre rischi propri, ad esempio attacchi contraddittori.

Qualsiasi sistema di IA ESPLICABILE *conterrà potenziali rischi: sia minacce sia opportunità.*

La PROPENSIONE *al rischio degli stakeholder è il livello di rischio e degli obiettivi generali che essi sono disposti ad accettare.*

Per una IA ESPLICABILE, *una strategia di gestione del rischio dovrà tenere conto dei* QUATTRO *principi.*

1° PRINCIPIO: *Una* ESPLICAZIONE *è necessaria per un'IA spiegabile, ma la spiegazione stessa introduce rischi positivi e negativi.*

Un potenziale risultato negativo di avere una spiegazione è l'esposizione di dettagli proprietari.

2° PRINCIPIO: *Le spiegazioni devono essere significative per il pubblico.*

Una spiegazione significativa può fornire una visione più approfondita del sistema, ma può esporre la proprietà intellettuale o le vulnerabilità del sistema esponendone il funzionamento interno.
Una spiegazione non significativa, invece, rischia di essere ignorata o non riconosciuta come spiegazione.

3° PRINCIPIO: *Per essere utile, una spiegazione non solo deve essere significativa ma deve anche essere accurata,* **terzo principio**.

Un rischio potenziale rilevante è comunemente noto come RISCHIO DI MODELLO: *potenziali esiti negativi derivati da un modello non valido o applicato in modo errato.*

Una Esplicazione imprecisa può portare a un'interpretazione errata o a un'incomprensione di come funziona il sistema o può arrivare a un risultato.
Ad esempio: nel riconoscimento facciale, un esaminatore del volto umano potrebbe ricevere informazioni da un algoritmo di IA su quali parti del viso sono utili. Una spiegazione accurata può aiutare l'esaminatore a valutare più accuratamente la coppia di facce, mentre una spiegazione imprecisa potrebbe portare a una decisione sbagliata.

Ad esempio: nel sistema giudiziario, gli algoritmi di IA sono stati utilizzati nelle decisioni, come se un imputato potesse essere nuovamente arrestato.

4° PRINCIPIO: *L'altra principale fonte di* RISCHIO DEL MODELLO *è l'utilizzo del modello in modo errato o oltre i suoi limiti di conoscenza.*

Le spiegazioni che descrivono i limiti di conoscenza del sistema possono fornire garanzie che il modello non stia funzionando al di fuori dell'ambito di applicazione e alimentare la fiducia.

5.5 - PANORAMICA DEGLI ALGORITMI IA ESPLICABILI

Seguendo altre fonti, organizziamo le spiegazioni in due grandi categorie:

1) *modelli auto-interpretabili e*
2) *spiegazioni post-hoc.*

I modelli AUTO-INTERPRETABILI *sono i modelli dell'algoritmo (o rappresentazione dell'algoritmo stesso) che può essere letto e interpretato direttamente da un essere umano.*

In questo caso il modello stesso è la spiegazione.

Le SPIEGAZIONI POST-HOC *sono spiegazioni, spesso generate da altri strumenti software, che descrivono, spiegano o modellano l'algoritmo per dare un'idea di come funziona l'algoritmo.*

Spiegazioni post-hoc spesso possono essere utilizzate su algoritmi senza alcuna conoscenza interna di come funziona l'algoritmo, a condizione che possa essere interrogato per output su input scelti.

1) MODELLI AUTO-INTERPRETABILI

I modelli AUTO-INTERPRETABILI *sono modelli che sono essi stessi le spiegazioni.*

Non solo spiegano l'intero modello a livello globale, ma esaminando ogni input attraverso il modello, la simulazione dell'input sul modello auto-interpretabile può fornire una spiegazione locale per ogni decisione.

Alcuni dei modelli auto-interpretabili più comuni includono alberi decisionali e modelli di regressione (inclusa la regressione logistica).

Questi modelli includono elenchi decisionali, set decisionali, prototipi (campioni rappresentativi di ciascuna classe) Kim, regole di combinazione di caratteristiche che classificano completamente insiemi di input Kuhn, elenchi di regole **bayesiane**, *alberi decisionali additivi e varianti migliorate degli alberi decisionali.*

Con i modelli auto-interpretabili, alcune fonti affermano un compromesso tra accuratezza e interpretabilità: i modelli AUTO-INTERPRETABILI *sono meno accurati dei modelli* POST-HOC *perché esiste un compromesso tra rendere il modello più esatto oppure più significativo.*

2) SPIEGAZIONI POST-HOC

Le SPIEGAZIONI POST-HOC *sono raggruppate in due tipi:*

1° LOCALI: *spiega un sottoinsieme di decisioni o è una spiegazione per decisione*,

2° GLOBALI: *produce un modello che approssima il modello non interpretabile*.

In alcuni casi, una GLOBALE *può anche fornire spiegazioni* LOCALI *simulandole su input specifici per fornire spiegazioni* LOCALI *per quei singoli input.*

Come semplici esempi, si consideri una regressione logistica (che potrebbe essere un modello auto-interpretabile o un'approssimazione post-hoc a un modello opaco).

I coefficienti di regressione forniscono una spiegazione GLOBALE *che spiega tutti gli input.*

Tuttavia, è possibile collegare l'input con i pesi e quindi utilizzare tali pesi per spiegare l'output dell'algoritmo.

1° TIPO: SPIEGAZIONI LOCALI

Le spiegazioni locali spiegano un sottoinsieme di input.

Il tipo più comune di spiegazione locale è una spiegazione per decisione o decisione singola, che fornisce una spiegazione per l'output dell'algoritmo o per la decisione su un singolo punto di input.

Un algoritmo di spiegazione locale comunemente usato è LIME (LOCAL INTERPRETABLE MODEL AGNOSTIC EXPLAINER). *Esso prende una decisione e, interrogando i punti vicini, costruisce un modello interpretabile che rappresenta la decisione locale, quindi utilizza quel modello per fornire spiegazioni per funzione.*

Il modello predefinito scelto è la regressione logistica.

Per le immagini, quindi interroga il modello con uno spazio di ricerca casuale in cui varia i super pixel omessi e sostituiti con tutto nero (o un colore a scelta dell'utente).

Un altro algoritmo di spiegazione locale comunemente usato è SHAP (SHAPLEY ADDITIVE EXPLANATIONS). *Esso fornisce un'importanza per funzione per un input su un problema di regressione convertendo lo scenario in un gioco di coalizione dalla teoria dei giochi e quindi producendo i valori di Shapley da quel gioco. Inoltre, tratta le caratteristiche come i giocatori, il valore delle caratteristiche rispetto a un valore predefinito come le strategie e l'output del sistema come il guadagno, formando un gioco di coalizione dall'input.*

Un'altra spiegazione locale comune è una CONTROFATTUALE.

Una CONTROFATTUALE *è una spiegazione che dice "se l'input fosse quest'altro nuovo input, il sistema avrebbe preso una decisione diversa".*

Un altro tipo popolare di spiegazioni locali per problemi sui dati dell'immagine sono i PIXEL DI SALIENZA.

I PIXEL DI SALIENZA *colorano ogni pixel a seconda di quanto quel pixel contribuisce alla decisione di classificazione.*

Uno dei primi algoritmi di salienza è CLASS ACTIVATION MAPS (CAM).

2° TIPO: SPIEGAZIONI GLOBALI

Le spiegazioni globali producono SPIEGAZIONI POST-HOC sull'intero algoritmo.

Spesso, ciò comporta la produzione di un modello GLOBALE *per un algoritmo o un sistema.*

Una spiegazione GLOBALE *è rappresentata dai* PDP (PARTIAL DEPENDENCE PLOTS).

Un grafico a dipendenza parziale mostra la variazione marginale della risposta prevista quando la caratteristica (valore di quella specifica colonna di dati o componente) cambia.

I PDP *sono utili per determinare se una relazione tra una caratteristica e la risposta è lineare o più complessa.*

Nelle reti neurali profonde, uno di questi algoritmi globali è TCAV *(*TESTING WITH CONCEPT ACTIVATION VECTORS*).*

TCAV *desidera spiegare una rete neurale in un modo più user-friendly, rappresentando lo stato della rete neurale come una ponderazione lineare di concetti umani, chiamati* CONCEPT ACTIVATION VECTORS (CAV).

Il TCAV è stato applicato per spiegare gli algoritmi di classificazione delle immagini attraverso l'apprendimento dei CAV, incluso il colore, per vedere come i colori hanno influenzato le decisioni del classificatore di immagini.

ATTACCHI CONTRO LA SPIEGABILITÀ

L'accuratezza della spiegazione (PRINCIPIO 3) è una componente importante delle spiegazioni.

A volte, se una spiegazione non ha un'accuratezza del 100%, può essere sfruttata da avversari che manipolano l'output di un classificatore su piccole perturbazioni di un input per nascondere i pregiudizi di un sistema.

5.6 - VALUTAZIONE DI ALGORITMI IA ESPLICABILI

VALUTARE IL SIGNIFICATO – FORWARD/COUNTERFACTUAL

Un modo per misurare il **SIGNIFICATO** di una spiegazione consiste nel misurare la capacità di simulare l'umano.

Questa è essenzialmente la capacità di una persona di comprendere un modello di apprendimento automatico nella misura in cui sarebbe in grado di prendere gli stessi dati di input del modello e di comprendere i parametri del modello in modo tale da poter produrre una previsione dal modello stesso in un ragionevole lasso di tempo.

La capacità di simulare il modello stesso rifletterebbe un alto grado di comprensione.

Questo è in genere misurato per i modelli auto-interpretabili come un modo per misurare la complessità del modello.

Hase e Bansal discutono di due tipi di simulazione umana: SIMULAZIONE IN AVANTI (FORWARD SIMULATION), che è quando un essere umano prevede l'output di un sistema per un dato input; SIMULAZIONE CONTROFATTUALE (COUNTERFACTUAL SIMULATION), quando a un essere umano viene dato un input e un output.

Devono prevedere quale output fornirebbe il sistema se l'input fosse cambiato in un modo particolare.

Un'altra strategia per valutare la significatività è chiedere agli esseri umani di completare un'attività utilizzando l'output del sistema fornito come input, quindi misurare il tempo impiegato dall'uomo e l'accuratezza della decisione sull'attività.

Lai e Tan hanno testato la significatività in un'attività di rilevamento di inganni. Il compito era determinare se le recensioni degli hotel fossero autentiche o ingannevoli.

L'accuratezza umana del rilevamento dell'inganno è stata confrontata quando è stata fornita solo la recensione stessa e quando loro sono state presentate le spiegazioni da una macchina.

Questo confronto consente di equiparare l'accuratezza delle decisioni umane con e senza assistenza/spiegazioni della macchina.

VALUTAZIONE DELL'ACCURATEZZA DELLA SPIEGAZIONE

L'accuratezza della spiegazione è strettamente correlata al lavoro sulla "FEDELTÀ".

Un modo in cui questo è stato testato è simulare modelli utilizzando l'output del sistema come verità di base e valutando le spiegazioni POST-HOC *utilizzando una metrica di apprendimento automatico.*

5.7 - GLI ESSERI UMANI COME GRUPPO DI CONFRONTO PER L'IA ESPLICABILE

Quando si considerano le prestazioni degli esseri umani e dei sistemi di IA, ci sono differenze di opinione abbastanza significative riguardo alle aspettative sulle prestazioni.

Alcuni sostengono che dovremmo mantenere le macchine a uno standard molto più elevato rispetto agli umani, mentre altri credono che sia sufficiente che le macchine siano semplicemente buone quanto gli umani.

Una cascata di domande interessanti e difficili sorgono da questo divario filosofico generale, come ad esempio:

- ✓ *Quanto devono essere le macchine migliori degli umani?*
- ✓ *In che modo devono essere migliori?*
- ✓ *Come misuriamo "buono come"?*

Indipendentemente da dove si cade in questo particolare dibattito filosofico, è comunque utile considerare le prestazioni umane come una linea di base.

Indipendentemente dall'IA, anche gli esseri umani che operano da soli prendono decisioni ad alto rischio con l'aspettativa che siano spiegabili. Ad esempio, medici, giudici, avvocati e forense. Ci si aspetta spesso che gli scienziati del settore forniscano una motivazione per i loro giudizi.

In che modo queste spiegazioni offerte aderiscono ai nostri quattro principi?

Ci siamo concentrati rigorosamente sulle spiegazioni umane dei propri giudizi e decisioni (ad esempio, "perché sei arrivato a questa conclusione o scelta?"),

Non su eventi esterni (ad esempio, "perché il cielo è blu?" o "perché un evento è verificarsi?").

*Gli eventi esterni accompagnati da spiegazioni possono essere utili per il ragionamento umano e per formulare previsioni. Ciò è coerente con il desiderio di un'*IA ESPLICABILE.

*Tuttavia, come delineato in quanto segue, le spiegazioni prodotte dall'uomo per i propri giudizi, decisioni e conclusioni sono in gran parte inaffidabili. Gli esseri umani come gruppo di confronto per l'*IA ESPLICABILE *possono informare lo sviluppo di metriche di riferimento e portare a una migliore comprensione delle dinamiche della collaborazione uomo-macchina.*

PRINCIPIO DELLA SPIEGAZIONE

Questo principio afferma che un sistema per essere considerato ESPLICABILE fornisca una spiegazione.

Gli esseri umani sono in grado di produrre una varietà di tipi di spiegazione. Tuttavia, la produzione di spiegazioni verbali può interferire con i processi decisionali e di ragionamento.

Si pensa che man mano che si acquisisce esperienza, i processi sottostanti diventino più automatici, al di fuori della consapevolezza cosciente, e quindi più difficili da spiegare verbalmente.

Ciò produce una tensione simile che esiste per l'IA stessa: si pensa spesso che il desiderio di un'elevata precisione derivi da riduzioni della spiegabilità.

Più in generale, i processi che si verificano con una consapevolezza cosciente limitata possono essere danneggiati richiedendo che la decisione stessa sia espressa in modo esplicito. Un esempio di questo viene dal rilevamento delle bugie.

Il rilevamento delle bugie basato sul giudizio esplicito se una persona sta dicendo o meno la verità o una bugia è in genere impreciso.

Tuttavia, quando i giudizi vengono forniti tramite attività di categorizzazione implicita, aggirando quindi un giudizio esplicito, l'accuratezza del rilevamento della menzogna può essere migliorata.

Questo suggerisce che l'individuazione della menzogna può essere un processo inconscio che viene interrotto quando costretto a diventare cosciente.

Insieme, questi risultati suggeriscono che alcune valutazioni degli esseri umani potrebbero essere più accurate se lasciate automatiche e implicite, rispetto a richiedere un giudizio o una spiegazione espliciti.

I giudizi umani e il processo decisionale possono spesso operare come una scatola chiusa e interferire con questo processo può essere dannoso per l'accuratezza di una decisione.

PRINCIPIO DEL SIGNIFICATO

Per soddisfare questo principio, il sistema fornisce spiegazioni comprensibili per il pubblico previsto.

Per questo, ci siamo concentrati sulla capacità degli esseri umani di interpretare il modo in cui un altro essere umano è arrivato ad una conclusione.

Questo significa:

1) *che il pubblico giunge alla stessa conclusione voluta dalla persona che fornisce la spiegazione, e*
2) *che il pubblico è d'accordo su quale sia la conclusione sulla base di una spiegazione.*

Un caso analogo all'IA spiegabile per l'interazione uomo-uomo è quello di uno scienziato forense che spiega prove forensi a persone comuni (ad esempio, membri di una giuria).

PRINCIPIO DELLA PRECISIONE DELLA SPIEGAZIONE

Questo principio afferma che la spiegazione di un sistema riflette correttamente le sue ragioni per generare un determinato output e/o riflette accuratamente il suo processo.

Per gli esseri umani, questo è analogo a una spiegazione dei propri processi decisionali che riflette veramente i processi mentali dietro quella decisione.

Per il tipo di introspezione correlata all'accuratezza della spiegazione, è ben documentato che, sebbene le persone spesso riferiscano il proprio ragionamento per le decisioni, ciò non riflette in modo affidabile un'introspezione accurata o significativa.

Questa è stata coniata "ILLUSIONE DELL'INTROSPEZIONE": termine per indicare che le informazioni ottenute guardando all'interno dei propri contenuti mentali si basano su nozioni errate che hanno valore.

Le persone inventano ragioni per le loro decisioni, anche quelle ritenute immutabili, come le opinioni personali.

In effetti, il ragionamento cosciente delle persone non sembra sempre verificarsi prima della decisione espressa. Invece, l'evidenza suggerisce che le persone prendano la loro decisione e quindi applichino le ragioni per tali decisioni dopo il fatto.

Dal punto di vista delle neuroscienze, i marcatori neurali di una decisione possono verificarsi fino a 10 secondi prima della consapevolezza cosciente di una persona.
Questa scoperta suggerisce che i processi decisionali iniziano molto prima della nostra consapevolezza cosciente.

Le persone sono in gran parte inconsapevoli della loro incapacità di introspezione in modo accurato.

*Ciò è documentato attraverso studi sulla "***cecità della scelta***" in cui le persone non ricordano accuratamente le loro decisioni precedenti.*

Nonostante questo ricordo impreciso, i partecipanti forniscono ragioni per effettuare selezioni che in realtà non hanno mai fatto.

Per gli studi che non coinvolgono la memoria a lungo termine, i partecipanti hanno anche dimostrato di non essere consapevoli dei modi in cui valutano i giudizi percettivi. Ad esempio, le persone sono imprecise quando segnalano quali caratteristiche facciali usano per determinare l'identità di qualcuno.

Sulla base della nostra definizione di ACCURATEZZA DELLA SPIEGAZIONE, *questi risultati non supportano l'idea che gli esseri umani soddisfino in modo affidabile questi criteri.*

Per numerosi compiti, gli esseri umani possono essere estremamente accurati ma non possono verbalizzare il loro processo decisionale.

PRINCIPIO DEI LIMITI DELLA CONOSCENZA

Questo principio afferma che il sistema funziona solo:

1. nelle condizioni in cui è stato progettato e
2. quando raggiunge una sufficiente fiducia nel proprio risultato o nelle proprie azioni.

Ci siamo concentrati sul fatto se gli esseri umani valutano correttamente le proprie capacità e accuratezza e se sanno quando segnalare che non conoscono una risposta.

Esistono diversi modi per verificare se le persone possono valutare i propri limiti di conoscenza.

- *Un metodo consiste nel chiedere ai partecipanti di prevedere quanto credono di aver eseguito bene o eseguiranno bene un compito, rispetto ad altri (ad esempio, in quale percentile cadranno i loro punteggi rispetto ad altri responsabili del compito).*

- *Un altro modo per testare la consapevolezza dei limiti di conoscenza è ottenere una misura della loro sicurezza di risposta, con una maggiore sicurezza che indica se una persona crede con maggiore probabilità di essere corretta.*

Come dimostrato dal noto effetto Dunning-Kruger, la maggior parte delle persone stima in modo impreciso le proprie capacità rispetto agli altri.

Una scoperta simile è che le persone, inclusi gli esperti, generalmente non predicono bene la propria accuratezza e capacità quando è chiesto loro di stimare esplicitamente le prestazioni.

Tuttavia, una recente replica dell'effetto Dunning-Kruger per la percezione del viso ha mostrato che, sebbene le persone non predicessero in modo affidabile la loro accuratezza, le loro stime delle capacità variavano di conseguenza con la difficoltà del compito.

Questo suggerisce che sebbene il valore esatto (ad es. Percentuale di prestazione prevista rispetto ad altri o percentuale di correttezza prevista) possa essere errato, le persone possono modulare la direzione della prestazione prevista in modo appropriato (ad es.: sapere che un compito era più o meno difficile per loro) .

Per una varietà di giudizi e decisioni, le persone spesso sanno quando hanno commesso errori, anche in assenza di feedback.

*Per usare la testimonianza oculare come esempio pertinente: sebbene la fiducia e l'accuratezza abbiano ripetutamente dimostrato di essere debolmente correlate, la fiducia di una persona predice la loro accuratezza in assenza di "*CONTAMINAZIONE*" attraverso il processo interrogatorio e il tempo prolungato tra l'evento e il momento del ricordo.*

Pertanto, le carenze umane nel valutare i propri limiti di conoscenza sono simili a quelle di produrre spiegazioni stesse.

Quando è chiesto esplicitamente di produrre una spiegazione, queste spiegazioni possono interferire con processi più automatici acquisiti dall'esperienza; spesso non riflettono accuratamente i veri processi cognitivi.

Allo stesso modo, quando alle persone è chiesto di prevedere o stimare esplicitamente il proprio livello di abilità rispetto agli altri, spesso sono imprecise.

Tuttavia, quando si richiede di valutare la loro fiducia per una determinata decisione rispetto a questo giudizio esplicito, le persone possono valutare la loro accuratezza a livelli superiori al caso.

Ciò suggerisce che le persone hanno una visione dei propri limiti di conoscenza, sebbene questa comprensione possa essere limitata o debole in alcuni casi.