NIST AI 100-1

Framework per la gestione del rischio dell'intelligenza artificiale (AI RMF 1.0)

Traduzione italiana a cura di Giuseppe Pelligra

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-1

Framework per la gestione del rischio dell'intelligenza artificiale (AI RMF 1.0)

Questa pubblicazione è disponibile gratuitamente
all'indirizzo:
https://doi.org/10.6028/NIST.AI.100-1

Gennaio 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Alcune entità commerciali, apparecchiature o materiali possono essere identificati in questo documento al fine di descrivere adeguatamente una procedura sperimentale o un concetto. Tale identificazione non intende implicare raccomandazione o approvazione da parte del National Institute of Standards and Technology, né intende implicare che tali entità, materiali o apparecchiature siano necessariamente i migliori disponibili per lo scopo.

Questa pubblicazione è disponibile gratuitamente all'indirizzo: https://doi.org/10.6028/NIST.AI.100-1

Crediti e avvertenza sulla traduzione

Il presente documento è una traduzione italiana non ufficiale del documento **NIST AI 100-1 - Artificial Intelligence Risk Management Framework (AI RMF 1.0)**, pubblicato dal **National Institute of Standards and Technology (NIST), U.S. Department of Commerce**, nel gennaio 2023.

Il documento originale è disponibile gratuitamente sul sito ufficiale NIST tramite DOI: **10.6028/NIST.AI.100-1**.

Questa traduzione è stata realizzata da **Giuseppe Pelligra** a scopo divulgativo, formativo e professionale, per favorire la comprensione in lingua italiana dei principi contenuti nel NIST AI Risk Management Framework.

Questa non è una traduzione ufficiale NIST. NIST non ha revisionato, approvato, sponsorizzato o certificato questa traduzione. In caso di dubbi interpretativi, discrepanze o utilizzo in contesti formali, normativi, contrattuali o consulenziali, deve essere sempre consultata la versione originale in lingua inglese pubblicata da NIST.

Eventuali marchi, loghi, riferimenti istituzionali e denominazioni presenti nel documento appartengono ai rispettivi titolari. L'utilizzo del nome NIST e dei riferimenti al documento originale ha esclusivamente finalità di attribuzione della fonte.

Versione traduzione italiana: 1.0 **Data:** giugno 2026

Programma di aggiornamento e versioni

L'Artificial Intelligence Risk Management Framework (AI RMF) è concepito come un documento vivo.

NIST riesaminerà regolarmente il contenuto e l'utilità del Framework per stabilire se sia opportuno aggiornarlo; una revisione con contributi formali della comunità AI è prevista non oltre il 2028. Il Framework utilizzerà un sistema di versionamento a due numeri per tracciare e identificare modifiche maggiori e minori. Il primo numero rappresenterà la generazione dell'AI RMF e dei documenti collegati, ad esempio 1.0, e cambierà solo in caso di revisioni maggiori. Le revisioni minori saranno tracciate usando ".n" dopo il numero di generazione, ad esempio 1.1. Tutte le modifiche saranno registrate in una tabella di controllo versione che identifica cronologia, numero di versione, data della modifica e descrizione. NIST prevede di aggiornare frequentemente l'AI RMF Playbook. I commenti sull'AI RMF Playbook possono essere inviati in qualsiasi momento via email a AIframework@nist.gov e saranno riesaminati e integrati con cadenza semestrale.

Indice

Elenco delle tabelle

Elenco delle figure

Sintesi esecutiva

Le tecnologie di intelligenza artificiale (AI) hanno un potenziale significativo per trasformare la società e la vita delle persone, dal commercio e dalla salute ai trasporti, alla cybersecurity, all'ambiente e al pianeta. Le tecnologie AI possono favorire una crescita economica inclusiva e sostenere progressi scientifici che migliorano le condizioni del nostro mondo. Tuttavia, le tecnologie AI pongono anche rischi che possono incidere negativamente su individui, gruppi, organizzazioni, comunità, società, ambiente e pianeta. Come i rischi di altri tipi di tecnologia, i rischi AI possono emergere in diversi modi ed essere caratterizzati come di lungo o breve termine, ad alta o bassa probabilità, sistemici o localizzati, e ad alto o basso impatto.

> L'AI RMF definisce un sistema AI come un sistema ingegnerizzato o basato su macchine che, per un determinato insieme di obiettivi, può generare output quali previsioni, raccomandazioni o decisioni che influenzano ambienti reali o virtuali. I sistemi AI sono progettati per operare con livelli variabili di autonomia (adattato da: OECD Recommendation on AI:2019; ISO / IEC 22989:2022).

Sebbene esistano numerosi standard e best practice per aiutare le organizzazioni a mitigare i rischi dei sistemi software tradizionali o basati su informazioni, i rischi posti dai sistemi AI sono per molti aspetti unici (si veda l'Appendice B). I sistemi AI, ad esempio, possono essere addestrati su dati che cambiano nel tempo, talvolta in modo significativo e inatteso, influenzando funzionalità e affidabilità del sistema in modi difficili da comprendere. I sistemi AI e i contesti in cui sono distribuiti sono spesso complessi, rendendo difficile rilevare e rispondere ai malfunzionamenti quando si verificano. I sistemi AI sono intrinsecamente socio-tecnici: sono influenzati dalle dinamiche sociali e dal comportamento umano. I rischi e i benefici AI possono emergere dall'interazione tra aspetti tecnici e fattori sociali relativi a come un sistema è usato, alle sue interazioni con altri sistemi AI, a chi lo gestisce e al contesto sociale in cui è distribuito.

Questi rischi rendono l'AI una tecnologia particolarmente sfidante da distribuire e utilizzare sia per le organizzazioni sia nella società. Senza controlli adeguati, i sistemi AI possono amplificare, perpetuare o aggravare risultati iniqui o indesiderati per individui e comunità. Con controlli adeguati, i sistemi AI possono mitigare e gestire risultati iniqui.

La gestione del rischio AI è una componente chiave dello sviluppo e dell'uso responsabile dei sistemi AI. Le pratiche di AI responsabile possono aiutare ad allineare le decisioni sulla progettazione, lo sviluppo e gli usi dei sistemi AI agli scopi e ai valori previsti. I concetti fondamentali dell'AI responsabile enfatizzano centralità umana, responsabilità sociale e sostenibilità. La gestione del rischio AI può guidare usi e pratiche responsabili inducendo organizzazioni e team interni che progettano, sviluppano e distribuiscono AI a riflettere in modo più critico sul contesto e sui potenziali impatti negativi e positivi, anche inattesi. Comprendere e gestire i rischi dei sistemi AI contribuirà ad accrescere l'affidabilità e, a sua volta, a coltivare la fiducia pubblica.

> La responsabilità sociale può riferirsi alla responsabilità dell'organizzazione "per gli impatti delle sue decisioni e attività sulla società e sull'ambiente mediante un comportamento trasparente ed etico" (ISO 26000:2010). La sostenibilità si riferisce allo "stato del sistema globale, inclusi gli aspetti ambientali, sociali ed economici, nel quale i bisogni del presente sono soddisfatti senza compromettere la capacità delle generazioni future di soddisfare i propri bisogni" (ISO / IEC TR 24368:2022). L'AI responsabile è intesa a produrre tecnologie che siano anche eque e soggette ad accountability. L'aspettativa è che le pratiche organizzative siano svolte in conformità alla "responsabilità professionale", definita da ISO come un approccio che "mira ad assicurare che i professionisti che progettano, sviluppano o distribuiscono sistemi e applicazioni AI o prodotti o sistemi basati su AI riconoscano la loro posizione unica nell'esercitare influenza su persone, società e futuro dell'AI" (ISO / IEC TR 24368:2022).

Come disposto dal National Artificial Intelligence Initiative Act del 2020 (P.L. 116-283), l'obiettivo dell'AI RMF è offrire una risorsa alle organizzazioni che progettano, sviluppano, distribuiscono o usano sistemi AI per aiutarle a gestire i molti rischi dell'AI e promuovere uno sviluppo e un uso dei sistemi AI affidabili e responsabili. Il Framework è pensato per essere volontario, rispettoso dei diritti, non specifico per settore e agnostico rispetto ai casi d'uso, offrendo flessibilità a organizzazioni di ogni dimensione e in tutti i settori e ambiti della società per implementare gli approcci contenuti nel Framework.

Il Framework è progettato per dotare organizzazioni e individui, qui denominati attori AI, di approcci che aumentino l'affidabilità dei sistemi AI e aiutino a favorire nel tempo la progettazione, lo sviluppo, la distribuzione e l'uso responsabile dei sistemi AI. Gli attori AI sono definiti dall'Organisation for Economic Co-operation and Development (OECD) come "coloro che svolgono un ruolo attivo nel ciclo di vita del sistema AI, incluse le organizzazioni e gli individui che distribuiscono o gestiscono AI" [OECD (2019) *Artificial Intelligence in Society* - OECD iLibrary] (si veda l'Appendice A).

L'AI RMF è pensato per essere pratico, per adattarsi al panorama dell'AI mentre le tecnologie AI continuano a svilupparsi, e per essere reso operativo dalle organizzazioni in gradi e capacità variabili, così che la società possa beneficiare dell'AI ed essere al tempo stesso protetta dai suoi potenziali danni.

Il Framework e le risorse di supporto saranno aggiornati, ampliati e migliorati sulla base dell'evoluzione tecnologica, del panorama degli standard a livello globale e dell'esperienza e del feedback della comunità AI. NIST continuerà ad allineare l'AI RMF e le linee guida correlate agli standard, alle linee guida e alle pratiche internazionali applicabili. Con l'uso dell'AI RMF emergeranno ulteriori lezioni utili a informare futuri aggiornamenti e risorse aggiuntive.

Il Framework è diviso in due parti. La Parte 1 discute come le organizzazioni possano inquadrare i rischi correlati all'AI e descrive i destinatari previsti. Successivamente vengono analizzati rischi AI e affidabilità, delineando le caratteristiche dei sistemi AI affidabili, tra cui validità e attendibilità, sicurezza, protezione e resilienza, accountability e trasparenza, spiegabilità e interpretabilità, privacy rafforzata, ed equità con gestione dei bias dannosi.

La Parte 2 comprende il "Core" del Framework. Descrive quattro funzioni specifiche per aiutare le organizzazioni ad affrontare nella pratica i rischi dei sistemi AI. Tali funzioni, GOVERN, MAP, MEASURE e MANAGE, sono ulteriormente suddivise in categorie e sottocategorie. Mentre GOVERN si applica a tutte le fasi dei processi e delle procedure di gestione del rischio AI delle organizzazioni, le funzioni MAP, MEASURE e MANAGE possono essere applicate in contesti specifici di sistema AI e in fasi specifiche del ciclo di vita AI.

Risorse aggiuntive relative al Framework sono incluse nell'AI RMF Playbook, disponibile tramite il sito NIST AI RMF:
https://www.nist.gov/itl/ai-risk-management-framework.

Lo sviluppo dell'AI RMF da parte di NIST in collaborazione con i settori privato e pubblico è disposto ed è coerente con le più ampie iniziative NIST sull'AI richieste dal National AI Initiative Act del 2020, dalle raccomandazioni della National Security Commission on Artificial Intelligence e dal *Plan for Federal Engagement in Developing Technical Standards and Related Tools*. Il coinvolgimento della comunità AI durante lo sviluppo di questo Framework, tramite risposte a una Request for Information formale, tre workshop molto partecipati, commenti pubblici su un concept paper e su due bozze del Framework, discussioni in molteplici forum pubblici e numerosi incontri di piccoli gruppi, ha informato lo sviluppo dell'AI RMF 1.0, nonché la ricerca, lo sviluppo e la valutazione AI condotti da NIST e da altri. La ricerca prioritaria e le ulteriori linee guida che miglioreranno questo Framework saranno raccolte in una AI Risk Management Framework Roadmap associata, alla quale NIST e la comunità più ampia potranno contribuire.

Parte 1: Informazioni fondamentali

1. Inquadrare il rischio

La gestione del rischio AI offre un percorso per minimizzare i potenziali impatti negativi dei sistemi AI, come minacce alle libertà e ai diritti civili, fornendo al contempo opportunità per massimizzare gli impatti positivi. Affrontare, documentare e gestire efficacemente i rischi AI e i potenziali impatti negativi può portare a sistemi AI più affidabili.

1.1 Comprendere e affrontare rischi, impatti e danni

Nel contesto dell'AI RMF, il rischio si riferisce alla misura composita della probabilità che un evento si verifichi e della magnitudine o grado delle conseguenze dell'evento corrispondente. Gli impatti, o conseguenze, dei sistemi AI possono essere positivi, negativi o entrambi e possono tradursi in opportunità o minacce (adattato da: ISO 31000:2018). Quando si considera l'impatto negativo di un potenziale evento, il rischio è funzione di 1) impatto negativo, o magnitudine del danno, che sorgerebbe se la circostanza o l'evento si verificasse, e 2) probabilità di accadimento (adattato da: OMB Circular A-130:2016). Impatti negativi o danni possono essere subiti da individui, gruppi, comunità, organizzazioni, società, ambiente e pianeta.

> "La gestione del rischio si riferisce ad attività coordinate per dirigere e controllare un'organizzazione con riferimento al rischio" (fonte: ISO 31000:2018).

Sebbene i processi di gestione del rischio affrontino in genere impatti negativi, questo Framework offre approcci per minimizzare gli impatti negativi previsti dei sistemi AI e identificare opportunità per massimizzare gli impatti positivi. Gestire efficacemente il rischio di potenziali danni potrebbe portare a sistemi AI più affidabili e liberare potenziali benefici per persone, organizzazioni e sistemi/ecosistemi. La gestione del rischio può consentire a sviluppatori e utenti AI di comprendere gli impatti e tenere conto delle limitazioni e incertezze intrinseche dei loro modelli e sistemi; ciò può migliorare le prestazioni complessive del sistema, la sua affidabilità e la probabilità che le tecnologie AI siano usate in modi benefici.

L'AI RMF è progettato per affrontare nuovi rischi quando emergono. Questa flessibilità è particolarmente importante quando gli impatti non sono facilmente prevedibili e le applicazioni evolvono. Sebbene alcuni rischi e benefici AI siano noti, valutare gli impatti negativi e il grado dei danni può essere difficile. La Figura 1 fornisce esempi di potenziali danni che possono essere correlati ai sistemi AI.

Le attività di gestione del rischio AI dovrebbero considerare che gli esseri umani possono presumere che i sistemi AI funzionino, e funzionino bene, in tutti i contesti. Ad esempio, correttamente o meno, i sistemi AI sono spesso percepiti come più oggettivi degli esseri umani o come dotati di capacità maggiori rispetto al software generale.

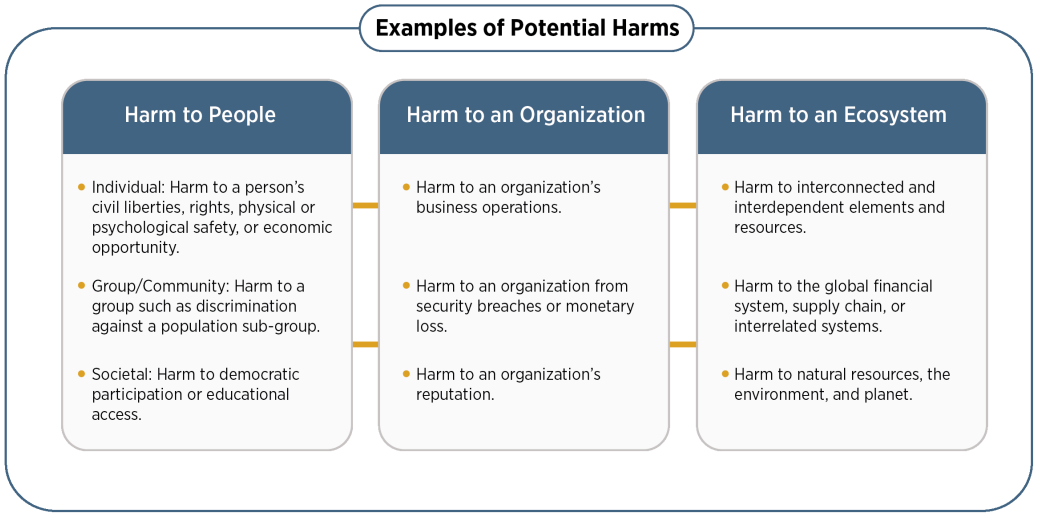

Fig. 1. Esempi di potenziali danni correlati ai sistemi AI. Sistemi AI affidabili e il loro uso responsabile possono mitigare i rischi negativi e contribuire a benefici per persone, organizzazioni ed ecosistemi.

1.2 Sfide per la gestione del rischio AI

Diverse sfide sono descritte di seguito. Dovrebbero essere considerate nella gestione dei rischi finalizzata all'affidabilità dell'AI.

1.2.1 Misurazione del rischio

I rischi o i malfunzionamenti AI che non sono ben definiti o adeguatamente compresi sono difficili da misurare in modo quantitativo o qualitativo. L'incapacità di misurare correttamente i rischi AI non implica necessariamente che un sistema AI ponga un rischio alto o basso. Alcune sfide di misurazione del rischio includono:

Rischi correlati a software, hardware e dati di terze parti: dati o sistemi di terze parti possono accelerare ricerca e sviluppo e facilitare la transizione tecnologica. Possono anche complicare la misurazione del rischio. Il rischio può emergere sia dai dati, dal software o dall'hardware di terze parti in sé, sia dal modo in cui sono usati. Metriche o metodologie di rischio usate dall'organizzazione che sviluppa il sistema AI potrebbero non allinearsi con quelle dell'organizzazione che distribuisce o gestisce il sistema. Inoltre, l'organizzazione che sviluppa il sistema AI potrebbe non essere trasparente sulle metriche o metodologie di rischio usate. Misurazione e gestione del rischio possono essere complicate dal modo in cui i clienti usano o integrano dati o sistemi di terze parti in prodotti o servizi AI, specialmente senza sufficienti strutture interne di governance e salvaguardie tecniche. In ogni caso, tutte le parti e gli attori AI dovrebbero gestire il rischio nei sistemi AI che sviluppano, distribuiscono o usano come componenti autonomi o integrati.

Tracciamento dei rischi emergenti: gli sforzi di gestione del rischio delle organizzazioni saranno rafforzati dall'identificazione e dal tracciamento dei rischi emergenti e dalla considerazione di tecniche per misurarli.

Gli approcci di valutazione dell'impatto dei sistemi AI possono aiutare gli attori AI a comprendere potenziali impatti o danni in contesti specifici.

Disponibilità di metriche attendibili: l'attuale mancanza di consenso su metodi di misurazione robusti e verificabili per rischio e affidabilità, e sulla loro applicabilità a diversi casi d'uso AI, è una sfida di misurazione del rischio AI. Potenziali insidie nel misurare il rischio negativo o i danni includono il fatto che lo sviluppo di metriche è spesso un'iniziativa istituzionale e può riflettere involontariamente fattori non correlati all'impatto sottostante. Inoltre, gli approcci di misurazione possono essere semplificati eccessivamente, manipolati, privi di sfumature critiche, usati in modi inattesi o incapaci di considerare differenze tra gruppi e contesti interessati.

Gli approcci per misurare gli impatti su una popolazione funzionano meglio se riconoscono che il contesto conta, che i danni possono incidere in modo diverso su gruppi o sottogruppi differenti e che comunità o altri sottogruppi potenzialmente danneggiati non sono sempre utenti diretti di un sistema.

Rischio in diverse fasi del ciclo di vita AI: misurare il rischio in una fase iniziale del ciclo di vita AI può produrre risultati diversi rispetto alla misurazione in una fase successiva; alcuni rischi possono essere latenti in un dato momento e crescere mentre i sistemi AI si adattano ed evolvono. Inoltre, attori AI diversi lungo il ciclo di vita possono avere prospettive di rischio differenti. Ad esempio, uno sviluppatore AI che rende disponibile software AI, come modelli pre-addestrati, può avere una prospettiva di rischio diversa da un attore AI responsabile della distribuzione di quel modello pre-addestrato in uno specifico caso d'uso. Tali deployer potrebbero non riconoscere che i loro usi specifici possono comportare rischi diversi da quelli percepiti dallo sviluppatore iniziale. Tutti gli attori AI coinvolti condividono responsabilità nella progettazione, nello sviluppo e nella distribuzione di un sistema AI affidabile e idoneo allo scopo.

Rischio in contesti reali: sebbene misurare i rischi AI in laboratorio o in un ambiente controllato possa fornire insight importanti prima della distribuzione, tali misure possono differire dai rischi che emergono in ambienti operativi reali.

Imperscrutabilità: i sistemi AI imperscrutabili possono complicare la misurazione del rischio. L'imperscrutabilità può derivare dalla natura opaca dei sistemi AI, dalla limitata spiegabilità o interpretabilità, dalla mancanza di trasparenza o documentazione nello sviluppo o nella distribuzione del sistema AI, oppure da incertezze intrinseche nei sistemi AI.

Baseline umana: la gestione del rischio dei sistemi AI destinati ad aumentare o sostituire attività umane, ad esempio il decision making, richiede una qualche forma di metriche di baseline per il confronto. Questo è difficile da sistematizzare, poiché i sistemi AI svolgono compiti diversi, e li svolgono diversamente, rispetto agli esseri umani.

1.2.2 Tolleranza al rischio

Sebbene l'AI RMF possa essere usato per prioritizzare il rischio, non prescrive la tolleranza al rischio. La tolleranza al rischio si riferisce alla disponibilità dell'organizzazione o dell'attore AI (si veda l'Appendice A) a sopportare il rischio per conseguire i propri obiettivi. La tolleranza al rischio può essere influenzata da requisiti legali o regolamentari (adattato da: ISO GUIDE 73). La tolleranza al rischio e il livello di rischio accettabile per organizzazioni o società sono fortemente contestuali e specifici per applicazione e caso d'uso. Le tolleranze al rischio possono essere influenzate da policy e norme stabilite da proprietari di sistemi AI, organizzazioni, settori, comunità o policy maker. È probabile che le tolleranze al rischio cambino nel tempo con l'evoluzione di sistemi AI, policy e norme. Organizzazioni diverse possono avere tolleranze al rischio diverse per via delle loro specifiche priorità organizzative e considerazioni sulle risorse.

Conoscenze e metodi emergenti per informare meglio i tradeoff danno/costo-beneficio continueranno a essere sviluppati e discussi da imprese, governi, mondo accademico e società civile. Nella misura in cui le sfide per specificare le tolleranze al rischio AI restano irrisolte, possono esistere contesti nei quali un framework di gestione del rischio non è ancora prontamente applicabile per mitigare rischi AI negativi.

> Il Framework è pensato per essere flessibile e per integrare pratiche di rischio esistenti, che dovrebbero allinearsi a leggi, regolamenti e norme applicabili. Le organizzazioni dovrebbero seguire regolamenti e linee guida esistenti per criteri, tolleranza e risposta al rischio stabiliti da requisiti organizzativi, di dominio, disciplina, settore o professione. Alcuni settori o industrie possono avere definizioni consolidate di danno o requisiti consolidati di documentazione, reporting e disclosure. All'interno dei settori, la gestione del rischio può dipendere da linee guida esistenti per applicazioni specifiche e impostazioni di casi d'uso. Dove non esistono linee guida consolidate, le organizzazioni dovrebbero definire una tolleranza al rischio ragionevole. Una volta definita la tolleranza, questo AI RMF può essere usato per gestire i rischi e documentare i processi di gestione del rischio.

1.2.3 Prioritizzazione del rischio

Tentare di eliminare del tutto il rischio negativo può essere controproducente nella pratica, perché non tutti gli incidenti e i malfunzionamenti possono essere eliminati. Aspettative irrealistiche sul rischio possono portare le organizzazioni ad allocare risorse in modo da rendere il triage del rischio inefficiente o impraticabile, oppure da sprecare risorse scarse. Una cultura di gestione del rischio può aiutare le organizzazioni a riconoscere che non tutti i rischi AI sono uguali e che le risorse possono essere allocate intenzionalmente. Sforzi di gestione del rischio azionabili delineano linee guida chiare per valutare l'affidabilità di ciascun sistema AI sviluppato o distribuito da un'organizzazione. Policy e risorse dovrebbero essere prioritizzate in base al livello di rischio valutato e al potenziale impatto di un sistema AI. Il grado in cui un sistema AI può essere personalizzato o adattato allo specifico contesto d'uso dal deployer AI può costituire un fattore contributivo.

Quando si applica l'AI RMF, i rischi che l'organizzazione determina come più elevati per i sistemi AI in un determinato contesto d'uso richiedono la prioritizzazione più urgente e il processo di gestione del rischio più approfondito. Nei casi in cui un sistema AI presenti livelli di rischio negativo inaccettabili, ad esempio quando impatti negativi significativi sono imminenti, danni gravi si stanno effettivamente verificando o sono presenti rischi catastrofici, sviluppo e distribuzione dovrebbero cessare in modo sicuro finché i rischi non possano essere gestiti sufficientemente. Se lo sviluppo, la distribuzione e i casi d'uso di un sistema AI risultano a basso rischio in un contesto specifico, ciò può suggerire una prioritizzazione potenzialmente inferiore.

La prioritizzazione del rischio può differire tra sistemi AI progettati o distribuiti per interagire direttamente con esseri umani e sistemi AI che non lo sono. Una prioritizzazione iniziale più elevata può essere richiesta in contesti in cui il sistema AI è addestrato su grandi dataset composti da dati sensibili o protetti, come informazioni di identificazione personale, o in cui gli output dei sistemi AI hanno impatto diretto o indiretto sugli esseri umani. Sistemi AI progettati per interagire solo con sistemi computazionali e addestrati su dataset non sensibili, ad esempio dati raccolti dall'ambiente fisico, possono richiedere una prioritizzazione iniziale inferiore. Tuttavia, valutare e prioritizzare regolarmente il rischio in base al contesto resta importante, poiché anche sistemi AI non rivolti agli esseri umani possono avere implicazioni a valle per la sicurezza o per la società.

Il rischio residuo, definito come rischio che rimane dopo il trattamento del rischio (fonte: ISO GUIDE 73), incide direttamente su utenti finali o individui e comunità interessati. Documentare i rischi residui richiederà al provider del sistema di considerare pienamente i rischi della distribuzione del prodotto AI e informerà gli utenti finali sui potenziali impatti negativi dell'interazione con il sistema.

1.2.4 Integrazione organizzativa e gestione del rischio

I rischi AI non dovrebbero essere considerati in isolamento. Attori AI diversi hanno responsabilità e consapevolezze diverse in funzione dei loro ruoli nel ciclo di vita. Ad esempio, le organizzazioni che sviluppano un sistema AI spesso non avranno informazioni su come il sistema possa essere usato. La gestione del rischio AI dovrebbe essere integrata e incorporata in strategie e processi più ampi di enterprise risk management. Trattare i rischi AI insieme ad altri rischi critici, come cybersecurity e privacy, produrrà un risultato più integrato ed efficienze organizzative.

L'AI RMF può essere utilizzato insieme a guide e framework correlati per gestire i rischi dei sistemi AI o rischi aziendali più ampi. Alcuni rischi correlati ai sistemi AI sono comuni ad altri tipi di sviluppo e distribuzione software. Esempi di rischi sovrapposti includono: preoccupazioni di privacy relative all'uso dei dati sottostanti per addestrare sistemi AI; implicazioni energetiche e ambientali associate a esigenze computazionali ad alta intensità di risorse; preoccupazioni di sicurezza relative a riservatezza, integrità e disponibilità del sistema e dei suoi dati di training e output; e sicurezza generale del software e dell'hardware sottostanti ai sistemi AI.

Le organizzazioni devono stabilire e mantenere meccanismi appropriati di accountability, ruoli e responsabilità, cultura e strutture di incentivo affinché la gestione del rischio sia efficace. L'uso del solo AI RMF non porterà a questi cambiamenti né fornirà gli incentivi appropriati. Una gestione del rischio efficace si realizza tramite impegno organizzativo ai livelli senior e può richiedere cambiamenti culturali in un'organizzazione o in un settore. Inoltre, organizzazioni di piccole e medie dimensioni che gestiscono rischi AI o implementano l'AI RMF possono affrontare sfide diverse rispetto alle grandi organizzazioni, a seconda delle loro capacità e risorse.

2. Destinatari

Identificare e gestire rischi AI e potenziali impatti, sia positivi sia negativi, richiede un ampio insieme di prospettive e attori lungo il ciclo di vita AI. Idealmente, gli attori AI rappresenteranno una diversità di esperienze, competenze e background e comprenderanno team diversificati dal punto di vista demografico e disciplinare. L'AI RMF è destinato a essere usato dagli attori AI lungo il ciclo di vita e le dimensioni dell'AI.

L'OECD ha sviluppato un framework per classificare le attività del ciclo di vita AI secondo cinque dimensioni socio-tecniche chiave, ciascuna con proprietà rilevanti per policy e governance AI, inclusa la gestione del rischio [OECD (2022) *OECD Framework for the Classification of AI systems* - OECD Digital Economy Papers]. La Figura 2 mostra queste dimensioni, leggermente modificate da NIST ai fini di questo framework. La modifica NIST evidenzia l'importanza dei processi di test, valutazione, verifica e validazione (TEVV) lungo l'intero ciclo di vita AI e generalizza il contesto operativo di un sistema AI.

Le dimensioni AI mostrate nella Figura 2 sono Application Context, Data and Input, AI Model e Task and Output. Gli attori AI coinvolti in queste dimensioni, che eseguono o gestiscono progettazione, sviluppo, distribuzione, valutazione e uso dei sistemi AI e guidano gli sforzi di gestione del rischio AI, costituiscono il pubblico primario dell'AI RMF.

Gli attori AI rappresentativi lungo le dimensioni del ciclo di vita sono elencati nella Figura 3 e descritti in dettaglio nell'Appendice A. Nell'AI RMF tutti gli attori AI collaborano per gestire i rischi e raggiungere gli obiettivi di AI affidabile e responsabile. Gli attori AI con competenze specifiche TEVV sono integrati lungo il ciclo di vita AI e hanno particolare probabilità di beneficiare del Framework. Svolti regolarmente, i compiti TEVV possono fornire insight relativi a standard o norme tecniche, sociali, legali ed etiche e possono aiutare ad anticipare impatti, valutare e tracciare rischi emergenti. Come processo regolare entro un ciclo di vita AI, il TEVV consente sia correzioni in corso d'opera sia gestione del rischio ex post.

La dimensione People & Planet al centro della Figura 2 rappresenta i diritti umani e il più ampio benessere della società e del pianeta. Gli attori AI in questa dimensione comprendono un pubblico separato dell'AI RMF che informa il pubblico primario. Questi attori AI possono includere associazioni di categoria, organizzazioni di sviluppo standard, ricercatori, gruppi di advocacy,

Fig. 2. Ciclo di vita e dimensioni chiave di un sistema AI. Adattato da OECD (2022) *OECD Framework for the Classification of AI systems* - OECD Digital Economy Papers. I due cerchi interni mostrano le dimensioni chiave dei sistemi AI e il cerchio esterno mostra le fasi del ciclo di vita AI. Idealmente, gli sforzi di gestione del rischio iniziano con la funzione Plan and Design nel contesto applicativo e sono svolti lungo l'intero ciclo di vita del sistema AI. Si veda la Figura 3 per gli attori AI rappresentativi.

gruppi ambientalisti, organizzazioni della società civile, utenti finali e individui e comunità potenzialmente impattati. Questi attori possono:

- assistere nel fornire contesto e comprensione di impatti potenziali ed effettivi;
- essere una fonte di norme e linee guida formali o quasi formali per la gestione del rischio AI;
- designare confini per il funzionamento dell'AI, tecnici, sociali, legali ed etici; e
- promuovere la discussione sui tradeoff necessari a bilanciare valori e priorità sociali relativi a libertà e diritti civili, equità, ambiente e pianeta, ed economia.

Una gestione del rischio efficace dipende da un senso di responsabilità collettiva tra gli attori AI mostrati nella Figura 3. Le funzioni dell'AI RMF, descritte nella Sezione 5, richiedono prospettive, discipline, professioni ed esperienze diverse. Team diversificati contribuiscono a una condivisione più aperta di idee e assunzioni sugli scopi e sulle funzioni della tecnologia, rendendo più espliciti questi aspetti impliciti. Questa prospettiva collettiva più ampia crea opportunità per far emergere problemi e identificare rischi esistenti ed emergenti.

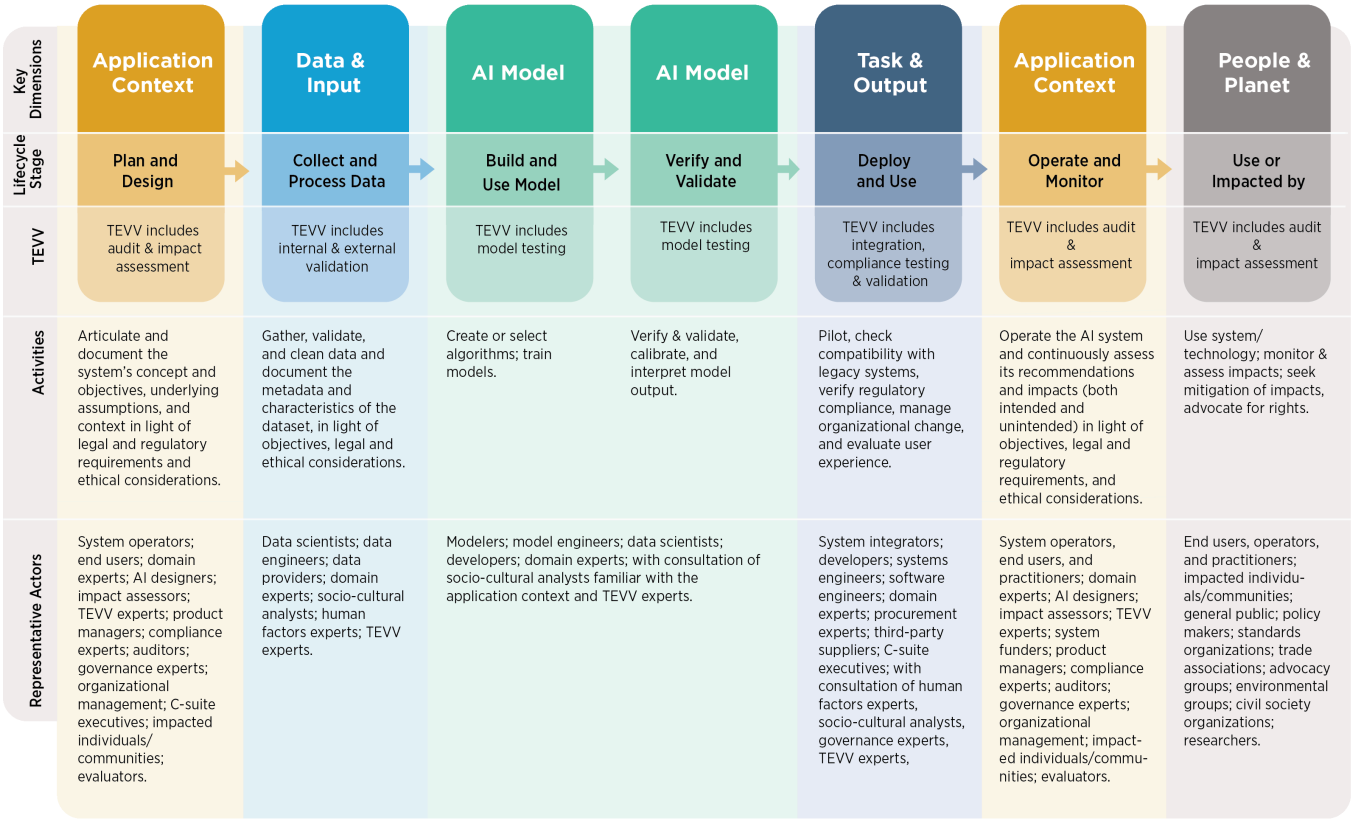

Fig. 3. Attori AI nelle fasi del ciclo di vita AI. Si veda l'Appendice A per descrizioni dettagliate dei compiti degli attori AI, inclusi i dettagli su test, valutazione, verifica e validazione. Gli attori AI nella dimensione AI Model (Figura 2) sono separati come buona pratica: chi costruisce e usa i modelli è distinto da chi verifica e valida i modelli.

3. Rischi AI e affidabilità

Per essere affidabili, i sistemi AI devono spesso rispondere a una molteplicità di criteri che hanno valore per le parti interessate. Approcci che rafforzano l'affidabilità dell'AI possono ridurre i rischi AI negativi. Questo Framework articola le seguenti caratteristiche dell'AI affidabile e offre indicazioni per affrontarle. Le caratteristiche dei sistemi AI affidabili includono: **validi e attendibili, sicuri, protetti e resilienti, con accountability e trasparenti, spiegabili e interpretabili, con privacy rafforzata, ed equi con bias dannosi gestiti.** Creare AI affidabile richiede di bilanciare ciascuna di queste caratteristiche in base al contesto d'uso del sistema AI. Sebbene tutte le caratteristiche siano attributi di sistema socio-tecnico, accountability e trasparenza riguardano anche processi e attività interni a un sistema AI e al suo ambiente esterno. Trascurare queste caratteristiche può aumentare la probabilità e la magnitudine di conseguenze negative.

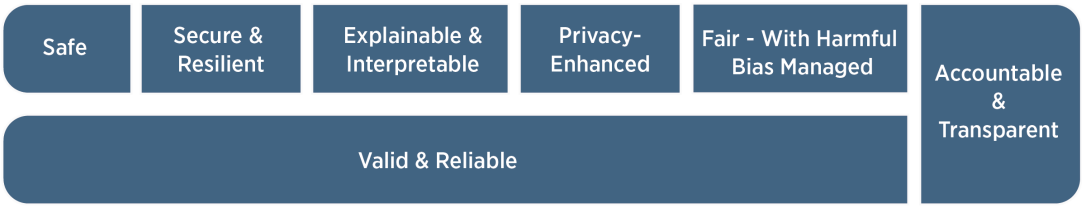

Fig. 4. Caratteristiche dei sistemi AI affidabili. Valido e attendibile è una condizione necessaria dell'affidabilità ed è mostrata come base delle altre caratteristiche. La caratteristica accountability e trasparenza è mostrata come riquadro verticale perché riguarda tutte le altre caratteristiche.

Le caratteristiche di affidabilità mostrate nella Figura 4 sono inestricabilmente legate al comportamento sociale e organizzativo, ai dataset usati dai sistemi AI, alla selezione dei modelli e algoritmi AI, alle decisioni di chi li costruisce e alle interazioni con gli esseri umani che forniscono insight e supervisione su tali sistemi. Il giudizio umano dovrebbe essere impiegato nel decidere le metriche specifiche relative alle caratteristiche di affidabilità dell'AI e i valori soglia precisi di tali metriche.

Affrontare individualmente le caratteristiche di affidabilità AI non assicurerà l'affidabilità del sistema AI; di solito sono coinvolti tradeoff, raramente tutte le caratteristiche si applicano in ogni contesto, e alcune saranno più o meno importanti in una determinata situazione. In ultima analisi, l'affidabilità è un concetto sociale che si distribuisce lungo uno spettro ed è forte solo quanto le sue caratteristiche più deboli.

Nel gestire i rischi AI, le organizzazioni possono affrontare decisioni difficili nel bilanciare queste caratteristiche. Ad esempio, in certi scenari possono emergere tradeoff tra ottimizzazione dell'interpretabilità e raggiungimento della privacy. In altri casi le organizzazioni possono affrontare un tradeoff tra accuratezza predittiva e interpretabilità. Oppure, in determinate condizioni come la scarsità di dati, tecniche di rafforzamento della privacy possono comportare una perdita di accuratezza, influenzando decisioni su equità e altri valori in certi domini.

Gestire i tradeoff richiede di considerare il contesto decisionale. Queste analisi possono evidenziare l'esistenza e l'estensione dei tradeoff tra misure differenti, ma non rispondono alla domanda su come navigare il tradeoff. Ciò dipende dai valori in gioco nel contesto rilevante e dovrebbe essere risolto in modo trasparente e adeguatamente giustificabile.

Esistono molteplici approcci per rafforzare la consapevolezza contestuale nel ciclo di vita AI. Ad esempio, esperti di dominio possono contribuire alla valutazione dei risultati TEVV e lavorare con team di prodotto e distribuzione per allineare i parametri TEVV ai requisiti e alle condizioni di distribuzione. Se adeguatamente dotato di risorse, aumentare ampiezza e diversità degli input da parti interessate e attori AI rilevanti lungo il ciclo di vita può migliorare le opportunità di informare valutazioni sensibili al contesto e identificare benefici e impatti positivi del sistema AI. Queste pratiche possono aumentare la probabilità che i rischi che sorgono nei contesti sociali siano gestiti in modo appropriato.

La comprensione e il trattamento delle caratteristiche di affidabilità dipendono dal ruolo specifico di un attore AI nel ciclo di vita AI. Per un dato sistema AI, un progettista o sviluppatore AI può avere una percezione delle caratteristiche diversa rispetto al deployer.

> Le caratteristiche di affidabilità spiegate in questo documento si influenzano reciprocamente. Sistemi altamente protetti ma iniqui, sistemi accurati ma opachi e non interpretabili, e sistemi inaccurati ma protetti, con privacy rafforzata e trasparenti sono tutti indesiderabili. Un approccio completo alla gestione del rischio richiede di bilanciare i tradeoff tra le caratteristiche di affidabilità. È responsabilità congiunta di tutti gli attori AI determinare se una tecnologia AI sia uno strumento appropriato o necessario per un dato contesto o scopo, e come usarla responsabilmente. La decisione di commissionare o distribuire un sistema AI dovrebbe basarsi su una valutazione contestuale delle caratteristiche di affidabilità e dei relativi rischi, impatti, costi e benefici, ed essere informata da un ampio insieme di parti interessate.

3.1 Valido e attendibile

La validazione è la "conferma, attraverso la fornitura di evidenza oggettiva, che i requisiti per uno specifico uso o applicazione previsti sono stati soddisfatti" (fonte: ISO 9000:2015). La distribuzione di sistemi AI inaccurati, non attendibili o scarsamente generalizzati a dati e contesti oltre il loro training crea e aumenta rischi AI negativi e riduce l'affidabilità.

L'attendibilità è definita nello stesso standard come "capacità di un elemento di funzionare come richiesto, senza guasti, per un determinato intervallo di tempo, in condizioni date" (fonte: ISO / IEC TS 5723:2022). L'attendibilità è un obiettivo per la correttezza complessiva del funzionamento del sistema AI nelle condizioni d'uso previste e per un dato periodo di tempo, compresa l'intera vita del sistema.

Accuratezza e robustezza contribuiscono alla validità e all'attendibilità dei sistemi AI e possono essere in tensione l'una con l'altra nei sistemi AI.

L'accuratezza è definita da ISO / IEC TS 5723:2022 come "prossimità dei risultati di osservazioni, computazioni o stime ai valori veri o ai valori accettati come veri". Le misure di accuratezza dovrebbero considerare metriche incentrate sulla computazione, ad esempio tassi di falsi positivi e falsi negativi, teaming umano-AI e dimostrare validità esterna, cioè generalizzabilità oltre le condizioni di training. Le misurazioni dell'accuratezza dovrebbero essere sempre abbinate a test set chiaramente definiti e realistici, rappresentativi delle condizioni d'uso previste, e a dettagli sulla metodologia di test; tali elementi dovrebbero essere inclusi nella documentazione associata. Le misurazioni dell'accuratezza possono includere la disaggregazione dei risultati per diversi segmenti di dati.

Robustezza o generalizzabilità è definita come "capacità di un sistema di mantenere il proprio livello di prestazione in una varietà di circostanze" (fonte: ISO / IEC TS 5723:2022). La robustezza è un obiettivo di funzionalità appropriata del sistema in un ampio insieme di condizioni e circostanze, inclusi usi dei sistemi AI non inizialmente anticipati. La robustezza richiede non solo che il sistema funzioni esattamente come negli usi attesi, ma anche che funzioni in modi che minimizzino potenziali danni alle persone se opera in un contesto inatteso.

Validità e attendibilità dei sistemi AI distribuiti sono spesso valutate tramite test o monitoraggio continuo che confermano che un sistema stia funzionando come previsto. La misurazione di validità, accuratezza, robustezza e attendibilità contribuisce all'affidabilità e dovrebbe considerare che certi tipi di guasto possono causare danni maggiori. Gli sforzi di gestione del rischio AI dovrebbero prioritizzare la minimizzazione dei potenziali impatti negativi e possono dover includere intervento umano nei casi in cui il sistema AI non possa rilevare o correggere errori.

3.2 Sicuro

I sistemi AI non dovrebbero, "in condizioni definite, portare a uno stato nel quale vita umana, salute, proprietà o ambiente siano messi in pericolo" (fonte: ISO / IEC TS 5723:2022). Il funzionamento sicuro dei sistemi AI è migliorato tramite:

- pratiche responsabili di progettazione, sviluppo e distribuzione;
- informazioni chiare ai deployer sull'uso responsabile del sistema;
- decision making responsabile da parte di deployer e utenti finali; e
- spiegazioni e documentazione dei rischi basate su evidenza empirica di incidenti.

Diversi tipi di rischi di sicurezza possono richiedere approcci di gestione del rischio AI adattati al contesto e alla gravità dei potenziali rischi presentati. Rischi di sicurezza che pongono potenziale rischio di lesioni gravi o morte richiedono la prioritizzazione più urgente e il processo di gestione del rischio più approfondito.

Integrare considerazioni di sicurezza durante il ciclo di vita, iniziando il prima possibile con pianificazione e progettazione, può prevenire guasti o condizioni che rendano un sistema pericoloso. Altri approcci pratici per la sicurezza AI spesso riguardano simulazione rigorosa e test nel dominio, monitoraggio in tempo reale e capacità di arrestare, modificare o prevedere intervento umano nei sistemi che deviano dalla funzionalità prevista o attesa.

Gli approcci di gestione del rischio per la sicurezza AI dovrebbero trarre spunto da sforzi e linee guida per la sicurezza in ambiti quali trasporti e sanità, e allinearsi a linee guida o standard esistenti specifici per settore o applicazione.

3.3 Protetto e resiliente

I sistemi AI, così come gli ecosistemi in cui sono distribuiti, possono essere detti resilienti se possono resistere a eventi avversi inattesi o cambiamenti inattesi nel loro ambiente o uso, oppure se possono mantenere funzioni e struttura di fronte a cambiamenti interni ed esterni e degradare in modo sicuro e ordinato quando necessario (adattato da: ISO / IEC TS 5723:2022). Preoccupazioni di sicurezza comuni riguardano esempi avversariali, data poisoning ed esfiltrazione di modelli, dati di training o altra proprietà intellettuale tramite endpoint del sistema AI. Sistemi AI che possono mantenere riservatezza, integrità e disponibilità tramite meccanismi di protezione che impediscono accesso e uso non autorizzati possono essere detti protetti. Le linee guida del NIST Cybersecurity Framework e del Risk Management Framework rientrano tra quelle applicabili in questo ambito.

Protezione e resilienza sono caratteristiche correlate ma distinte. Mentre la resilienza è la capacità di tornare alla normale funzione dopo un evento avverso inatteso, la protezione include la resilienza ma comprende anche protocolli per evitare, proteggere, rispondere o recuperare da attacchi. La resilienza si collega alla robustezza e va oltre la provenienza dei dati per includere uso inatteso o avversariale, abuso o uso improprio del modello o dei dati.

3.4 Accountability e trasparenza

L'AI affidabile dipende dall'accountability. L'accountability presuppone trasparenza. La trasparenza riflette la misura in cui informazioni su un sistema AI e sui suoi output sono disponibili agli individui che interagiscono con tale sistema, indipendentemente dal fatto che siano consapevoli di farlo. Una trasparenza significativa fornisce accesso a livelli appropriati di informazione in base alla fase del ciclo di vita AI e adattati al ruolo o alla conoscenza degli attori AI o degli individui che interagiscono con il sistema AI o lo usano. Promuovendo livelli più elevati di comprensione, la trasparenza aumenta la fiducia nel sistema AI.

L'ambito di questa caratteristica si estende dalle decisioni di progettazione e dai dati di training al training del modello, alla struttura del modello, ai casi d'uso previsti e a come e quando sono state prese decisioni di distribuzione, post-distribuzione o dell'utente finale e da chi. La trasparenza è spesso necessaria per un ricorso azionabile relativo a output del sistema AI errati o che altrimenti conducono a impatti negativi. La trasparenza dovrebbe considerare l'interazione umano-AI: ad esempio, come un operatore o utente umano viene notificato quando viene rilevato un esito avverso potenziale o effettivo causato da un sistema AI.

Un sistema trasparente non è necessariamente un sistema accurato, con privacy rafforzata, protetto o equo. Tuttavia, è difficile determinare se un sistema opaco possieda tali caratteristiche, e farlo nel tempo mentre sistemi complessi evolvono.

Il ruolo degli attori AI dovrebbe essere considerato quando si cerca accountability per gli esiti dei sistemi AI. La relazione tra rischio e accountability associata all'AI e ai sistemi tecnologici più in generale differisce tra contesti culturali, legali, settoriali e sociali. Quando le conseguenze sono gravi, come quando sono in gioco vita e libertà, sviluppatori e deployer AI dovrebbero considerare un adeguamento proporzionato e proattivo delle proprie pratiche di trasparenza e accountability. Mantenere pratiche organizzative e strutture di governance per la riduzione del danno, come la gestione del rischio, può contribuire a sistemi con maggiore accountability.

Le misure per rafforzare trasparenza e accountability dovrebbero anche considerare l'impatto di tali sforzi sull'entità che li implementa, incluso il livello di risorse necessario e la necessità di salvaguardare informazioni proprietarie.

Mantenere la provenienza dei dati di training e supportare l'attribuzione delle decisioni del sistema AI a sottoinsiemi di dati di training può assistere sia trasparenza sia accountability. I dati di training possono inoltre essere soggetti a copyright e dovrebbero rispettare le leggi applicabili sui diritti di proprietà intellettuale.

Con l'evoluzione degli strumenti di trasparenza per sistemi AI e della relativa documentazione, gli sviluppatori di sistemi AI sono incoraggiati a testare diversi tipi di strumenti di trasparenza in cooperazione con i deployer AI per assicurare che i sistemi AI siano usati come previsto.

3.5 Spiegabile e interpretabile

La spiegabilità si riferisce a una rappresentazione dei meccanismi alla base del funzionamento dei sistemi AI, mentre l'interpretabilità si riferisce al significato dell'output dei sistemi AI nel contesto degli scopi funzionali per i quali sono progettati. Insieme, spiegabilità e interpretabilità aiutano chi gestisce o supervisiona un sistema AI, così come gli utenti di un sistema AI, a ottenere insight più profondi sulla funzionalità e sull'affidabilità del sistema, inclusi i suoi output. L'assunzione sottostante è che le percezioni di rischio negativo derivino dalla mancanza di capacità di dare senso o contestualizzare correttamente l'output del sistema. Sistemi AI spiegabili e interpretabili offrono informazioni che aiuteranno gli utenti finali a comprendere scopi e potenziale impatto di un sistema AI.

Il rischio derivante dalla mancanza di spiegabilità può essere gestito descrivendo come funzionano i sistemi AI, con descrizioni adattate a differenze individuali quali ruolo, conoscenze e livello di competenza dell'utente. Sistemi spiegabili possono essere sottoposti a debug e monitorati più facilmente e si prestano a documentazione, audit e governance più approfonditi.

I rischi per l'interpretabilità possono spesso essere affrontati comunicando una descrizione del motivo per cui un sistema AI ha formulato una determinata previsione o raccomandazione. Si vedano "Four Principles of Explainable Artificial Intelligence" e "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence".

Trasparenza, spiegabilità e interpretabilità sono caratteristiche distinte che si supportano reciprocamente. La trasparenza può rispondere alla domanda "che cosa è accaduto" nel sistema. La spiegabilità può rispondere alla domanda "come" una decisione è stata presa nel sistema. L'interpretabilità può rispondere alla domanda "perché" una decisione è stata presa dal sistema e quale sia il suo significato o contesto per l'utente.

3.6 Privacy rafforzata

La privacy si riferisce generalmente alle norme e pratiche che aiutano a salvaguardare autonomia, identità e dignità umane. Tali norme e pratiche affrontano tipicamente libertà dall'intrusione, limitazione dell'osservazione o capacità degli individui di acconsentire alla disclosure o controllare aspetti della propria identità, ad esempio corpo, dati, reputazione. Si veda *The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management*.

Valori di privacy come anonimato, riservatezza e controllo dovrebbero in generale guidare le scelte di progettazione, sviluppo e distribuzione dei sistemi AI. I rischi relativi alla privacy possono influenzare sicurezza, bias e trasparenza e comportare tradeoff con queste altre caratteristiche. Come sicurezza e protezione, specifiche caratteristiche tecniche di un sistema AI possono promuovere o ridurre la privacy. I sistemi AI possono anche presentare nuovi rischi per la privacy consentendo inferenze per identificare individui o informazioni precedentemente private su individui.

Le tecnologie di rafforzamento della privacy ("PETs") per l'AI, così come metodi di minimizzazione dei dati quali de-identificazione e aggregazione per determinati output di modello, possono supportare la progettazione di sistemi AI con privacy rafforzata. In certe condizioni, come la scarsità di dati, tecniche di rafforzamento della privacy possono comportare una perdita di accuratezza, influenzando decisioni su equità e altri valori in alcuni domini.

3.7 Equo, con bias dannosi gestiti

L'equità nell'AI include preoccupazioni per uguaglianza ed equity affrontando questioni come bias dannosi e discriminazione. Gli standard di equità possono essere complessi e difficili da definire perché le percezioni di equità differiscono tra culture e possono cambiare a seconda dell'applicazione. Gli sforzi di gestione del rischio delle organizzazioni saranno rafforzati dal riconoscere e considerare tali differenze. I sistemi in cui i bias dannosi sono mitigati non sono necessariamente equi. Ad esempio, sistemi in cui le previsioni sono in qualche modo bilanciate tra gruppi demografici possono comunque essere inaccessibili a persone con disabilità o interessate dal digital divide, oppure possono aggravare disparità esistenti o bias sistemici.

Il bias è più ampio del bilanciamento demografico e della rappresentatività dei dati. NIST ha identificato tre principali categorie di bias AI da considerare e gestire: sistemico, computazionale e statistico, e cognitivo umano. Ciascuna può verificarsi in assenza di pregiudizio, parzialità o intento discriminatorio. Il bias sistemico può essere presente nei dataset AI, nelle norme, pratiche e processi organizzativi lungo il ciclo di vita AI e nella società più ampia che usa sistemi AI. I bias computazionali e statistici possono essere presenti nei dataset AI e nei processi algoritmici, e spesso derivano da errori sistematici dovuti a campioni non rappresentativi. I bias cognitivi umani riguardano come un individuo o gruppo percepisce le informazioni del sistema AI per prendere una decisione o colmare informazioni mancanti, o come gli esseri umani pensano a scopi e funzioni di un sistema AI. I bias cognitivi umani sono onnipresenti nei processi decisionali lungo il ciclo di vita AI e nell'uso del sistema, inclusi progettazione, implementazione, funzionamento e manutenzione dell'AI.

Il bias esiste in molte forme e può radicarsi nei sistemi automatizzati che aiutano a prendere decisioni sulle nostre vite. Sebbene il bias non sia sempre un fenomeno negativo, i sistemi AI possono potenzialmente aumentare velocità e scala dei bias e perpetuare e amplificare danni a individui, gruppi, comunità, organizzazioni e società. Il bias è strettamente associato ai concetti di trasparenza e di equità nella società. Per ulteriori informazioni sul bias, incluse le tre categorie, si veda NIST Special Publication 1270, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*.

4. Efficacia dell'AI RMF

Le valutazioni dell'efficacia dell'AI RMF, inclusi i modi per misurare miglioramenti sostanziali nell'affidabilità dei sistemi AI, faranno parte delle future attività NIST, in collaborazione con la comunità AI.

Le organizzazioni e altri utenti del Framework sono incoraggiati a valutare periodicamente se l'AI RMF abbia migliorato la loro capacità di gestire i rischi AI, includendo, ma non limitandosi a, policy, processi, pratiche, piani di implementazione, indicatori, misurazioni e risultati attesi. NIST intende lavorare in modo collaborativo con altri per sviluppare metriche, metodologie e obiettivi per valutare l'efficacia dell'AI RMF e condividere ampiamente risultati e informazioni di supporto. Gli utenti del Framework dovrebbero beneficiare di:

- processi rafforzati per applicare le funzioni GOVERN, MAP, MEASURE e MANAGE al rischio AI e documentare chiaramente i risultati;
- maggiore consapevolezza delle relazioni e dei tradeoff tra caratteristiche di affidabilità, approcci socio-tecnici e rischi AI;
- processi espliciti per prendere decisioni go/no-go di commissionamento e distribuzione del sistema;
- policy, processi, pratiche e procedure consolidate per migliorare gli sforzi organizzativi di accountability relativi ai rischi dei sistemi AI;
- cultura organizzativa rafforzata che prioritizza identificazione e gestione dei rischi dei sistemi AI e dei potenziali impatti su individui, comunità, organizzazioni e società;
- migliore condivisione di informazioni all'interno e tra organizzazioni su rischi, processi decisionali, responsabilità, insidie comuni, pratiche TEVV e approcci al miglioramento continuo;
- maggiore conoscenza contestuale per aumentare la consapevolezza dei rischi a valle;
- coinvolgimento rafforzato di parti interessate e attori AI rilevanti; e
- capacità aumentata di TEVV dei sistemi AI e dei rischi associati.

Parte 2: Core e Profili

5. AI RMF Core

L'AI RMF Core fornisce risultati e azioni che abilitano dialogo, comprensione e attività per gestire i rischi AI e sviluppare responsabilmente sistemi AI affidabili. Come illustrato nella Figura 5, il Core è composto da quattro funzioni: **GOVERN**, **MAP**, **MEASURE** e **MANAGE**. Ciascuna di queste funzioni di alto livello è suddivisa in categorie e sottocategorie. Categorie e sottocategorie sono a loro volta suddivise in azioni e risultati specifici. Le azioni non costituiscono una checklist, né sono necessariamente un insieme ordinato di passaggi.

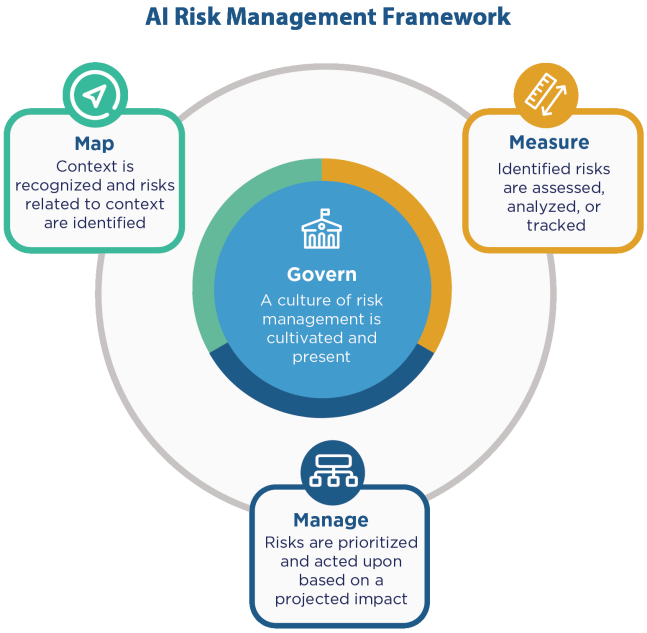

Fig. 5. Le funzioni organizzano al livello più alto le attività di gestione del rischio AI nelle funzioni GOVERN, MAP, MEASURE e MANAGE. La funzione GOVERN è concepita come funzione trasversale per informare ed essere integrata nelle altre tre funzioni.

La gestione del rischio dovrebbe essere continua, tempestiva e svolta lungo le dimensioni del ciclo di vita del sistema AI. Le funzioni dell'AI RMF Core dovrebbero essere eseguite in modo da riflettere prospettive diverse e multidisciplinari, includendo potenzialmente le opinioni di attori AI esterni all'organizzazione. Avere un team diversificato contribuisce a una condivisione più aperta di idee e assunzioni sugli scopi e sulle funzioni della tecnologia progettata, sviluppata, distribuita o valutata, creando opportunità per far emergere problemi e identificare rischi esistenti ed emergenti.

Una risorsa online di accompagnamento all'AI RMF, il NIST AI RMF Playbook, è disponibile per aiutare le organizzazioni a navigare l'AI RMF e raggiungerne i risultati tramite azioni tattiche suggerite che possono applicare nei propri contesti. Come l'AI RMF, il Playbook è volontario e le organizzazioni possono utilizzare i suggerimenti secondo necessità e interessi. Gli utenti del Playbook possono creare guide personalizzate selezionate dal materiale suggerito per il proprio uso e contribuire con suggerimenti da condividere con la comunità più ampia. Insieme all'AI RMF, il Playbook fa parte del NIST Trustworthy and Responsible AI Resource Center.

> Gli utenti del Framework possono applicare queste funzioni nel modo più adatto alle loro esigenze di gestione dei rischi AI, in base a risorse e capacità. Alcune organizzazioni possono scegliere tra categorie e sottocategorie; altre possono scegliere, e avere la capacità, di applicare tutte le categorie e sottocategorie. Presupposta l'esistenza di una struttura di governance, le funzioni possono essere svolte in qualsiasi ordine lungo il ciclo di vita AI quando ritenuto utile dall'utente del framework. Dopo aver istituito i risultati in GOVERN, la maggior parte degli utenti dell'AI RMF inizierebbe con la funzione MAP e proseguirebbe verso MEASURE o MANAGE. Qualunque sia il modo in cui gli utenti integrano le funzioni, il processo dovrebbe essere iterativo, con richiami incrociati tra funzioni quando necessario. Analogamente, vi sono categorie e sottocategorie con elementi che si applicano a più funzioni o che logicamente dovrebbero avvenire prima di determinate decisioni di sottocategoria.

5.1 GOVERN

La funzione GOVERN:

- coltiva e implementa una cultura di gestione del rischio nelle organizzazioni che progettano, sviluppano, distribuiscono, valutano o acquisiscono sistemi AI;
- delinea processi, documenti e schemi organizzativi che anticipano, identificano e gestiscono i rischi che un sistema può porre, anche a utenti e ad altri nella società, e procedure per raggiungere tali risultati;
- incorpora processi per valutare potenziali impatti;
- fornisce una struttura con cui le funzioni di gestione del rischio AI possono allinearsi a principi, policy e priorità strategiche organizzative;
- collega aspetti tecnici di progettazione e sviluppo del sistema AI a valori e principi organizzativi e abilita pratiche e competenze organizzative per gli individui coinvolti nell'acquisizione, training, distribuzione e monitoraggio di tali sistemi; e
- affronta l'intero ciclo di vita del prodotto e i processi associati, inclusi aspetti legali e altri temi relativi all'uso di sistemi software o hardware e dati di terze parti.

GOVERN è una funzione trasversale integrata in tutta la gestione del rischio AI e abilita le altre funzioni del processo. Aspetti di GOVERN, specialmente quelli relativi a compliance o valutazione, dovrebbero essere integrati in ciascuna delle altre funzioni. L'attenzione alla governance è un requisito continuo e intrinseco per una gestione efficace del rischio AI lungo la vita di un sistema AI e la gerarchia dell'organizzazione.

Una governance forte può guidare e rafforzare pratiche e norme interne per facilitare una cultura organizzativa del rischio. Le autorità di governance possono determinare le policy generali che dirigono missione, obiettivi, valori, cultura e tolleranza al rischio di un'organizzazione. La leadership senior stabilisce il tono della gestione del rischio nell'organizzazione e, con esso, la cultura organizzativa. Il management allinea gli aspetti tecnici della gestione del rischio AI a policy e operazioni. La documentazione può rafforzare la trasparenza, migliorare i processi di revisione umana e sostenere l'accountability nei team dei sistemi AI.

Dopo aver messo in atto strutture, sistemi, processi e team descritti nella funzione GOVERN, le organizzazioni dovrebbero beneficiare di una cultura orientata allo scopo e focalizzata sulla comprensione e gestione del rischio. Spetta agli utenti del Framework continuare a eseguire la funzione GOVERN mentre conoscenze, culture, bisogni o aspettative degli attori AI evolvono nel tempo.

Le pratiche relative alla governance dei rischi AI sono descritte nel NIST AI RMF Playbook. La Tabella 1 elenca categorie e sottocategorie della funzione GOVERN.

Tabella 1: Categorie e sottocategorie della funzione GOVERN.

Categorie	Sottocategorie
GOVERN 1: policy, processi, procedure e pratiche nell'organizzazione relative a mappatura, misurazione e gestione dei rischi AI sono presenti, trasparenti e implementati efficacemente.	**GOVERN 1.1:** requisiti legali e regolamentari che coinvolgono l'AI sono compresi, gestiti e documentati. **GOVERN 1.2:** le caratteristiche dell'AI affidabile sono integrate in policy, processi, procedure e pratiche organizzative. **GOVERN 1.3:** processi, procedure e pratiche sono presenti per determinare il livello necessario di attività di gestione del rischio in base alla tolleranza al rischio dell'organizzazione. **GOVERN 1.4:** il processo di gestione del rischio e i suoi risultati sono stabiliti tramite policy, procedure e altri controlli trasparenti basati sulle priorità di rischio organizzative.

Continua nella pagina successiva

Tabella 1: Categorie e sottocategorie della funzione GOVERN. (Segue)

Categorie	Sottocategorie
GOVERN 1 (segue)	**GOVERN 1.5:** monitoraggio continuo e revisione periodica del processo di gestione del rischio e dei suoi risultati sono pianificati e ruoli e responsabilità organizzativi chiaramente definiti, inclusa la frequenza della revisione periodica. **GOVERN 1.6:** sono presenti meccanismi per inventariare i sistemi AI e sono dotati di risorse secondo le priorità di rischio organizzative. **GOVERN 1.7:** sono presenti processi e procedure per dismettere e ritirare gradualmente i sistemi AI in modo sicuro e senza aumentare i rischi o diminuire l'affidabilità dell'organizzazione.
GOVERN 2: sono presenti strutture di accountability affinché team e individui appropriati siano abilitati, responsabili e formati per mappare, misurare e gestire i rischi AI.	**GOVERN 2.1:** ruoli, responsabilità e linee di comunicazione relative a mappatura, misurazione e gestione dei rischi AI sono documentati e chiari a individui e team in tutta l'organizzazione. **GOVERN 2.2:** personale e partner dell'organizzazione ricevono formazione sulla gestione del rischio AI per svolgere compiti e responsabilità coerentemente con policy, procedure e accordi correlati. **GOVERN 2.3:** la leadership esecutiva dell'organizzazione assume responsabilità per le decisioni sui rischi associati a sviluppo e distribuzione dei sistemi AI.
GOVERN 3: processi di diversità, equità, inclusione e accessibilità della forza lavoro sono prioritizzati nella mappatura, misurazione e gestione dei rischi AI lungo il ciclo di vita.	**GOVERN 3.1:** il decision making relativo a mappatura, misurazione e gestione dei rischi AI lungo il ciclo di vita è informato da un team diversificato, ad esempio per demografia, discipline, esperienza, competenza e background. **GOVERN 3.2:** sono presenti policy e procedure per definire e differenziare ruoli e responsabilità per configurazioni umano-AI e supervisione dei sistemi AI.
GOVERN 4: i team organizzativi sono impegnati in una cultura che considera e comunica il rischio AI.	**GOVERN 4.1:** policy e pratiche organizzative sono presenti per promuovere mentalità critica e orientata anzitutto alla sicurezza nella progettazione, sviluppo, distribuzione e uso dei sistemi AI, al fine di minimizzare potenziali impatti negativi.

Continua nella pagina successiva

Tabella 1: Categorie e sottocategorie della funzione GOVERN. (Segue)

Categorie	Sottocategorie
GOVERN 4 (segue)	**GOVERN 4.2:** i team organizzativi documentano i rischi e i potenziali impatti della tecnologia AI che progettano, sviluppano, distribuiscono, valutano e usano, e comunicano più ampiamente tali impatti. **GOVERN 4.3:** sono presenti pratiche organizzative per abilitare test AI, identificazione degli incidenti e condivisione delle informazioni.
GOVERN 5: sono presenti processi per un coinvolgimento robusto degli attori AI rilevanti.	**GOVERN 5.1:** policy e pratiche organizzative sono presenti per raccogliere, considerare, prioritizzare e integrare feedback da soggetti esterni al team che ha sviluppato o distribuito il sistema AI riguardo ai potenziali impatti individuali e sociali correlati ai rischi AI. **GOVERN 5.2:** sono stabiliti meccanismi per consentire al team che ha sviluppato o distribuito sistemi AI di incorporare regolarmente nel design e nell'implementazione del sistema feedback valutati provenienti da attori AI rilevanti.
GOVERN 6: policy e procedure sono presenti per affrontare rischi e benefici AI derivanti da software e dati di terze parti e da altri temi di supply chain.	**GOVERN 6.1:** policy e procedure sono presenti per affrontare rischi AI associati a entità terze, inclusi rischi di violazione della proprietà intellettuale o di altri diritti di terzi. **GOVERN 6.2:** sono presenti processi di contingency per gestire malfunzionamenti o incidenti in dati o sistemi AI di terze parti ritenuti ad alto rischio.

5.2 MAP

La funzione MAP stabilisce il contesto per inquadrare i rischi correlati a un sistema AI. Il ciclo di vita AI consiste in molte attività interdipendenti che coinvolgono un insieme diversificato di attori (si veda la Figura 3). Nella pratica, gli attori AI responsabili di una parte del processo spesso non hanno piena visibilità o controllo sulle altre parti e sui relativi contesti. Le interdipendenze tra queste attività e tra gli attori AI rilevanti possono rendere difficile anticipare in modo affidabile gli impatti dei sistemi AI. Ad esempio, decisioni iniziali nell'identificazione di scopi e obiettivi di un sistema AI possono alterarne comportamento e capacità, e le dinamiche dell'ambiente di distribuzione, come utenti finali o individui impattati, possono modellare gli impatti delle decisioni del sistema AI. Di conseguenza, le migliori intenzioni in una dimensione del ciclo di vita AI possono essere compromesse dalle interazioni con decisioni e condizioni in altre attività successive.

Questa complessità e i diversi livelli di visibilità possono introdurre incertezza nelle pratiche di gestione del rischio. Anticipare, valutare e altrimenti affrontare potenziali fonti di rischio negativo può mitigare tale incertezza e rafforzare l'integrità del processo decisionale.

Le informazioni raccolte durante l'esecuzione della funzione MAP abilitano la prevenzione del rischio negativo e informano decisioni per processi quali gestione del modello, nonché una decisione iniziale sull'appropriatezza o sulla necessità di una soluzione AI. I risultati della funzione MAP sono la base per le funzioni MEASURE e MANAGE. Senza conoscenza contestuale e consapevolezza dei rischi nei contesti identificati, la gestione del rischio è difficile da svolgere. La funzione MAP mira a rafforzare la capacità di un'organizzazione di identificare rischi e fattori contributivi più ampi.

L'implementazione di questa funzione è rafforzata dall'incorporazione di prospettive provenienti da un team interno diversificato e dal coinvolgimento di soggetti esterni al team che ha sviluppato o distribuito il sistema AI. Il coinvolgimento di collaboratori esterni, utenti finali, comunità potenzialmente impattate e altri può variare in base al livello di rischio di uno specifico sistema AI, alla composizione del team interno e alle policy organizzative. Raccogliere tali prospettive ampie può aiutare le organizzazioni a prevenire proattivamente rischi negativi e a sviluppare sistemi AI più affidabili mediante:

- miglioramento della capacità di comprendere i contesti;
- verifica delle assunzioni sul contesto d'uso;
- riconoscimento di quando i sistemi non sono funzionali entro o fuori dal contesto previsto;
- identificazione di usi positivi e benefici dei sistemi AI esistenti;
- migliore comprensione delle limitazioni nei processi AI e ML;
- identificazione di vincoli nelle applicazioni reali che possono portare a impatti negativi;
- identificazione di impatti negativi noti e prevedibili correlati all'uso previsto dei sistemi AI; e
- anticipazione dei rischi dell'uso dei sistemi AI oltre l'uso previsto.

Dopo aver completato la funzione MAP, gli utenti del Framework dovrebbero avere sufficiente conoscenza contestuale sugli impatti del sistema AI per informare una decisione iniziale go/no-go sull'opportunità di progettare, sviluppare o distribuire un sistema AI. Se si decide di procedere, le organizzazioni dovrebbero utilizzare le funzioni MEASURE e MANAGE insieme a policy e procedure messe in atto nella funzione GOVERN per assistere gli sforzi di gestione del rischio AI. Spetta agli utenti del Framework continuare ad applicare la funzione MAP ai sistemi AI mentre contesto, capacità, rischi, benefici e potenziali impatti evolvono nel tempo.

Le pratiche relative alla mappatura dei rischi AI sono descritte nel NIST AI RMF Playbook. La Tabella 2 elenca categorie e sottocategorie della funzione MAP.

Tabella 2: Categorie e sottocategorie della funzione MAP.

Categorie	Sottocategorie
MAP 1: il contesto è stabilito e compreso.	**MAP 1.1:** scopi previsti, usi potenzialmente benefici, leggi, norme e aspettative specifiche del contesto e ambienti prospettici in cui il sistema AI sarà distribuito sono compresi e documentati. Le considerazioni includono utenti specifici e aspettative, potenziali impatti positivi e negativi su individui, comunità, organizzazioni, società e pianeta, assunzioni e limitazioni relative a scopi, usi e rischi lungo il ciclo di vita di sviluppo o prodotto AI, e relative metriche TEVV e di sistema. **MAP 1.2:** attori AI interdisciplinari, competenze, skill e capacità per stabilire il contesto riflettono diversità demografica e ampia esperienza di dominio e utente; la loro partecipazione è documentata. Le opportunità di collaborazione interdisciplinare sono prioritizzate. **MAP 1.3:** la missione dell'organizzazione e gli obiettivi rilevanti per la tecnologia AI sono compresi e documentati. **MAP 1.4:** il valore di business o il contesto d'uso business è chiaramente definito o, nel caso di valutazione di sistemi AI esistenti, rivalutato. **MAP 1.5:** le tolleranze al rischio organizzative sono determinate e documentate. **MAP 1.6:** i requisiti di sistema, ad esempio "il sistema deve rispettare la privacy dei suoi utenti", sono raccolti presso gli attori AI rilevanti e compresi da essi. Le decisioni di design considerano implicazioni socio-tecniche per affrontare i rischi AI.
MAP 2: è svolta la categorizzazione del sistema AI.	**MAP 2.1:** sono definiti i compiti specifici e i metodi usati per implementare i compiti che il sistema AI supporterà, ad esempio classificatori, modelli generativi, recommender. **MAP 2.2:** sono documentate informazioni sui limiti di conoscenza del sistema AI e su come l'output del sistema possa essere utilizzato e supervisionato dagli esseri umani. La documentazione fornisce informazioni sufficienti ad assistere gli attori AI rilevanti nelle decisioni e azioni successive.

Continua nella pagina successiva

Tabella 2: Categorie e sottocategorie della funzione MAP. (Segue)

Categorie	Sottocategorie
MAP 2 (segue)	**MAP 2.3:** considerazioni di integrità scientifica e TEVV sono identificate e documentate, incluse quelle relative a disegno sperimentale, raccolta e selezione dei dati, ad esempio disponibilità, rappresentatività, idoneità, affidabilità del sistema e validazione del costrutto.
MAP 3: capacità AI, uso mirato, obiettivi e benefici e costi attesi confrontati con benchmark appropriati sono compresi.	**MAP 3.1:** sono esaminati e documentati i potenziali benefici della funzionalità e delle prestazioni previste del sistema AI. **MAP 3.2:** sono esaminati e documentati i costi potenziali, inclusi costi non monetari, risultanti da errori AI attesi o realizzati oppure da funzionalità e affidabilità del sistema, collegati alla tolleranza al rischio organizzativa. **MAP 3.3:** l'ambito applicativo mirato è specificato e documentato in base alla capacità del sistema, al contesto stabilito e alla categorizzazione del sistema AI. **MAP 3.4:** processi per competenza di operatori e practitioner rispetto a prestazioni e affidabilità del sistema AI, e standard tecnici e certificazioni rilevanti, sono definiti, valutati e documentati. **MAP 3.5:** processi per la supervisione umana sono definiti, valutati e documentati in conformità alle policy organizzative della funzione GOVERN.
MAP 4: rischi e benefici sono mappati per tutti i componenti del sistema AI, inclusi software e dati di terze parti.	**MAP 4.1:** sono presenti, seguiti e documentati approcci per mappare rischi della tecnologia AI e rischi legali dei suoi componenti, incluso l'uso di dati o software di terze parti, così come rischi di violazione della proprietà intellettuale o di altri diritti di terzi. **MAP 4.2:** controlli interni di rischio per componenti del sistema AI, incluse tecnologie AI di terze parti, sono identificati e documentati.
MAP 5: gli impatti su individui, gruppi, comunità, organizzazioni e società sono caratterizzati.	**MAP 5.1:** probabilità e magnitudine di ciascun impatto identificato, sia potenzialmente benefico sia dannoso, basate su uso atteso, usi passati di sistemi AI in contesti simili, report pubblici di incidenti, feedback da soggetti esterni al team che ha sviluppato o distribuito il sistema AI o altri dati, sono identificate e documentate.

Continua nella pagina successiva

Tabella 2: Categorie e sottocategorie della funzione MAP. (Segue)

Categorie	Sottocategorie
MAP 5 (segue)	**MAP 5.2:** pratiche e personale per supportare il coinvolgimento regolare di attori AI rilevanti e integrare feedback su impatti positivi, negativi e non anticipati sono presenti e documentati.

5.3 MEASURE

La funzione MEASURE impiega strumenti, tecniche e metodologie quantitative, qualitative o a metodi misti per analizzare, valutare, confrontare con benchmark e monitorare rischio AI e impatti correlati. Usa conoscenze rilevanti per i rischi AI identificati nella funzione MAP e informa la funzione MANAGE. I sistemi AI dovrebbero essere testati prima della distribuzione e regolarmente durante il funzionamento. Le misurazioni del rischio AI includono la documentazione di aspetti della funzionalità e dell'affidabilità dei sistemi.

Misurare i rischi AI include il tracciamento di metriche per caratteristiche di affidabilità, impatto sociale e configurazioni umano-AI. I processi sviluppati o adottati nella funzione MEASURE dovrebbero includere metodologie rigorose di testing software e valutazione delle prestazioni, con misure associate di incertezza, confronti con benchmark prestazionali e reporting e documentazione formalizzati dei risultati. Processi di revisione indipendente possono migliorare l'efficacia del testing e mitigare bias interni e potenziali conflitti di interesse.

Quando emergono tradeoff tra caratteristiche di affidabilità, la misurazione fornisce una base tracciabile per informare le decisioni di gestione. Le opzioni possono includere ricalibrazione, mitigazione dell'impatto o rimozione del sistema da progettazione, sviluppo, produzione o uso, oltre a una gamma di controlli compensativi, detective, deterrenti, direttivi e di recovery.

Dopo aver completato la funzione MEASURE, sono presenti, seguiti e documentati processi di test, valutazione, verifica e validazione (TEVV) oggettivi, ripetibili o scalabili, inclusi metriche, metodi e metodologie. Metriche e metodologie di misurazione dovrebbero aderire a norme scientifiche, legali ed etiche ed essere svolte in un processo aperto e trasparente. Può essere necessario sviluppare nuovi tipi di misurazione, qualitativa e quantitativa. Dovrebbe essere considerato il grado in cui ciascun tipo di misurazione fornisce informazioni uniche e significative alla valutazione dei rischi AI. Gli utenti del Framework rafforzeranno la propria capacità di valutare complessivamente l'affidabilità del sistema, identificare e tracciare rischi esistenti ed emergenti e verificare l'efficacia delle metriche. I risultati di misurazione saranno usati nella funzione MANAGE per assistere gli sforzi di monitoraggio e risposta al rischio. Spetta agli utenti del Framework continuare ad applicare la funzione MEASURE ai sistemi AI mentre conoscenze, metodologie, rischi e impatti evolvono nel tempo.

Le pratiche relative alla misurazione dei rischi AI sono descritte nel NIST AI RMF Playbook. La Tabella 3 elenca categorie e sottocategorie della funzione MEASURE.

Tabella 3: Categorie e sottocategorie della funzione MEASURE.

Categorie	Sottocategorie
MEASURE 1: metodi e metriche appropriati sono identificati e applicati.	**MEASURE 1.1:** approcci e metriche per la misurazione dei rischi AI enumerati durante la funzione MAP sono selezionati per l'implementazione iniziando dai rischi AI più significativi. I rischi o le caratteristiche di affidabilità che non saranno, o non potranno essere, misurati sono documentati correttamente. **MEASURE 1.2:** l'appropriatezza delle metriche AI e l'efficacia dei controlli esistenti sono valutate e aggiornate regolarmente, inclusi report di errori e potenziali impatti sulle comunità interessate. **MEASURE 1.3:** esperti interni che non hanno svolto il ruolo di sviluppatori front-line del sistema e/o assessor indipendenti sono coinvolti in valutazioni e aggiornamenti regolari. Esperti di dominio, utenti, attori AI esterni al team che ha sviluppato o distribuito il sistema AI e comunità interessate sono consultati a supporto delle valutazioni quando necessario secondo la tolleranza al rischio organizzativa.
MEASURE 2: i sistemi AI sono valutati per caratteristiche di affidabilità.	**MEASURE 2.1:** test set, metriche e dettagli sugli strumenti usati durante il TEVV sono documentati. **MEASURE 2.2:** valutazioni che coinvolgono soggetti umani soddisfano i requisiti applicabili, inclusa la protezione dei soggetti umani, e sono rappresentative della popolazione rilevante. **MEASURE 2.3:** criteri di prestazione o assurance del sistema AI sono misurati qualitativamente o quantitativamente e dimostrati per condizioni simili agli ambienti di distribuzione. Le misure sono documentate. **MEASURE 2.4:** funzionalità e comportamento del sistema AI e dei suoi componenti, come identificati nella funzione MAP, sono monitorati quando in produzione. **MEASURE 2.5:** il sistema AI da distribuire è dimostrato valido e attendibile. Le limitazioni della generalizzabilità oltre le condizioni in cui la tecnologia è stata sviluppata sono documentate.

Continua nella pagina successiva

Tabella 3: Categorie e sottocategorie della funzione MEASURE. (Segue)

Categorie	Sottocategorie
MEASURE 2 (segue)	**MEASURE 2.6:** il sistema AI è valutato regolarmente per rischi di sicurezza come identificati nella funzione MAP. Il sistema AI da distribuire è dimostrato sicuro, il suo rischio negativo residuo non supera la tolleranza al rischio e può fallire in modo sicuro, in particolare se fatto operare oltre i suoi limiti di conoscenza. Le metriche di sicurezza riflettono attendibilità e robustezza del sistema, monitoraggio in tempo reale e tempi di risposta ai guasti del sistema AI. **MEASURE 2.7:** protezione e resilienza del sistema AI, come identificate nella funzione MAP, sono valutate e documentate. **MEASURE 2.8:** rischi associati a trasparenza e accountability, come identificati nella funzione MAP, sono esaminati e documentati. **MEASURE 2.9:** il modello AI è spiegato, validato e documentato, e l'output del sistema AI è interpretato nel suo contesto, come identificato nella funzione MAP, per informare uso e governance responsabili. **MEASURE 2.10:** il rischio privacy del sistema AI, come identificato nella funzione MAP, è esaminato e documentato. **MEASURE 2.11:** equità e bias, come identificati nella funzione MAP, sono valutati e i risultati sono documentati. **MEASURE 2.12:** impatto ambientale e sostenibilità delle attività di training e gestione del modello AI, come identificati nella funzione MAP, sono valutati e documentati. **MEASURE 2.13:** efficacia delle metriche e dei processi TEVV impiegati nella funzione MEASURE è valutata e documentata.
MEASURE 3: sono presenti meccanismi per tracciare nel tempo i rischi AI identificati.	**MEASURE 3.1:** approcci, personale e documentazione sono presenti per identificare e tracciare regolarmente rischi AI esistenti, non anticipati ed emergenti sulla base di fattori quali prestazione prevista ed effettiva nei contesti distribuiti. **MEASURE 3.2:** approcci di tracciamento del rischio sono considerati per contesti in cui i rischi AI sono difficili da valutare usando tecniche di misurazione attualmente disponibili o in cui le metriche non sono ancora disponibili.

Continua nella pagina successiva

Tabella 3: Categorie e sottocategorie della funzione MEASURE. (Segue)

Categorie	Sottocategorie
MEASURE 3 (segue)	**MEASURE 3.3:** processi di feedback per utenti finali e comunità impattate per segnalare problemi e appellare gli esiti del sistema sono stabiliti e integrati nelle metriche di valutazione del sistema AI.
MEASURE 4: feedback sull'efficacia della misurazione è raccolto e valutato.	**MEASURE 4.1:** gli approcci di misurazione per identificare rischi AI sono collegati ai contesti di distribuzione e informati tramite consultazione con esperti di dominio e altri utenti finali. Gli approcci sono documentati. **MEASURE 4.2:** i risultati di misurazione riguardanti l'affidabilità del sistema AI nei contesti di distribuzione e lungo il ciclo di vita AI sono informati da input di esperti di dominio e attori AI rilevanti per validare se il sistema funzioni coerentemente come previsto. I risultati sono documentati. **MEASURE 4.3:** miglioramenti o declini misurabili delle prestazioni basati su consultazioni con attori AI rilevanti, incluse comunità interessate, e dati sul campo relativi a rischi e caratteristiche di affidabilità pertinenti al contesto sono identificati e documentati.

5.4 MANAGE

La funzione MANAGE comporta l'allocazione regolare delle risorse di rischio ai rischi mappati e misurati, come definito dalla funzione GOVERN. Il trattamento del rischio comprende piani per rispondere, recuperare e comunicare su incidenti o eventi.

Le informazioni contestuali ricavate dalla consultazione di esperti e dagli input di attori AI rilevanti, stabilite in GOVERN e svolte in MAP, sono utilizzate in questa funzione per diminuire la probabilità di guasti del sistema e impatti negativi. Le pratiche sistematiche di documentazione stabilite in GOVERN e utilizzate in MAP e MEASURE rafforzano gli sforzi di gestione del rischio AI e aumentano trasparenza e accountability. Sono presenti processi per valutare rischi emergenti, insieme a meccanismi di miglioramento continuo.

Dopo aver completato la funzione MANAGE, saranno presenti piani per prioritizzare il rischio e per monitoraggio e miglioramento regolari. Gli utenti del Framework avranno una capacità rafforzata di gestire i rischi dei sistemi AI distribuiti e di allocare risorse di gestione del rischio in base a rischi valutati e prioritizzati. Spetta agli utenti del Framework continuare ad applicare la funzione MANAGE ai sistemi AI distribuiti mentre metodi, contesti, rischi e bisogni o aspettative degli attori AI rilevanti evolvono nel tempo.

Le pratiche relative alla gestione dei rischi AI sono descritte nel NIST AI RMF Playbook. La Tabella 4 elenca categorie e sottocategorie della funzione MANAGE.

Tabella 4: Categorie e sottocategorie della funzione MANAGE.

Categorie	Sottocategorie
MANAGE 1: i rischi AI basati su assessment e altri output analitici delle funzioni MAP e MEASURE sono prioritizzati, affrontati e gestiti.	**MANAGE 1.1:** viene determinato se il sistema AI raggiunga gli scopi previsti e gli obiettivi dichiarati e se sviluppo o distribuzione debbano procedere. **MANAGE 1.2:** il trattamento dei rischi AI documentati è prioritizzato in base a impatto, probabilità e risorse o metodi disponibili. **MANAGE 1.3:** le risposte ai rischi AI ritenuti ad alta priorità, come identificati dalla funzione MAP, sono sviluppate, pianificate e documentate. Le opzioni di risposta al rischio possono includere mitigare, trasferire, evitare o accettare. **MANAGE 1.4:** i rischi negativi residui, definiti come somma di tutti i rischi non mitigati, per acquirenti a valle di sistemi AI e utenti finali sono documentati.
MANAGE 2: strategie per massimizzare benefici AI e minimizzare impatti negativi sono pianificate, preparate, implementate, documentate e informate da input di attori AI rilevanti.	**MANAGE 2.1:** le risorse richieste per gestire i rischi AI sono considerate, insieme a sistemi, approcci o metodi alternativi non AI praticabili, per ridurre magnitudine o probabilità dei potenziali impatti. **MANAGE 2.2:** meccanismi sono presenti e applicati per sostenere il valore dei sistemi AI distribuiti. **MANAGE 2.3:** sono seguite procedure per rispondere e recuperare da un rischio precedentemente sconosciuto quando viene identificato. **MANAGE 2.4:** meccanismi sono presenti e applicati, e responsabilità assegnate e comprese, per sostituire, disconnettere o disattivare sistemi AI che dimostrano prestazioni o esiti incoerenti con l'uso previsto.
MANAGE 3: rischi e benefici AI da entità terze sono gestiti.	**MANAGE 3.1:** rischi e benefici AI da risorse di terze parti sono monitorati regolarmente e i controlli di rischio sono applicati e documentati. **MANAGE 3.2:** modelli pre-addestrati usati per lo sviluppo sono monitorati come parte del monitoraggio e della manutenzione regolari del sistema AI.

Continua nella pagina successiva

Tabella 4: Categorie e sottocategorie della funzione MANAGE. (Segue)

Categorie	Sottocategorie
MANAGE 4: trattamenti del rischio, inclusi piani di risposta, recovery e comunicazione per i rischi AI identificati e misurati, sono documentati e monitorati regolarmente.	**MANAGE 4.1:** sono implementati piani di monitoraggio post-distribuzione del sistema AI, inclusi meccanismi per catturare e valutare input da utenti e altri attori AI rilevanti, appello e override, dismissione, risposta agli incidenti, recovery e change management. **MANAGE 4.2:** attività misurabili per miglioramenti continui sono integrate negli aggiornamenti del sistema AI e includono coinvolgimento regolare delle parti interessate, inclusi attori AI rilevanti. **MANAGE 4.3:** incidenti ed errori sono comunicati agli attori AI rilevanti, incluse comunità interessate. I processi per tracciare, rispondere e recuperare da incidenti ed errori sono seguiti e documentati.

6. Profili AI RMF

I profili di caso d'uso AI RMF sono implementazioni di funzioni, categorie e sottocategorie dell'AI RMF per uno specifico contesto o applicazione, basate su requisiti, tolleranza al rischio e risorse dell'utente del Framework: ad esempio, un profilo AI RMF per l'assunzione o un profilo AI RMF per l'equità abitativa. I profili possono illustrare e offrire insight su come il rischio possa essere gestito in varie fasi del ciclo di vita AI o in applicazioni specifiche di settore, tecnologia o uso finale. I profili AI RMF assistono le organizzazioni nel decidere come gestire al meglio il rischio AI in modo ben allineato ai propri obiettivi, considerando requisiti legali/regolamentari e best practice, e riflettendo priorità di gestione del rischio.

I profili temporali AI RMF sono descrizioni dello stato attuale o dello stato target desiderato di specifiche attività di gestione del rischio AI entro un dato settore, industria, organizzazione o contesto applicativo. Un AI RMF Current Profile indica come l'AI è attualmente gestita e i rischi correlati in termini di risultati attuali. Un Target Profile indica i risultati necessari per raggiungere gli obiettivi desiderati o target di gestione del rischio AI.

Il confronto tra Current e Target Profile probabilmente rivela gap da affrontare per soddisfare gli obiettivi di gestione del rischio AI. Possono essere sviluppati action plan per affrontare tali gap e realizzare risultati in una data categoria o sottocategoria. La prioritizzazione della mitigazione dei gap è guidata dai bisogni dell'utente e dai processi di gestione del rischio. Questo approccio basato sul rischio consente inoltre agli utenti del Framework di confrontare i propri approcci con altri approcci e valutare le risorse necessarie, ad esempio personale e finanziamenti, per raggiungere gli obiettivi di gestione del rischio AI in modo costo-efficace e prioritizzato.

I profili AI RMF cross-sector coprono i rischi di modelli o applicazioni che possono essere usati attraverso casi d'uso o settori. I profili cross-sector possono anche coprire come applicare GOVERN, MAP, MEASURE e MANAGE ai rischi per attività o processi di business comuni tra settori, come l'uso di large language model, servizi cloud-based o acquisizioni.

Questo Framework non prescrive template di profilo, consentendo flessibilità nell'implementazione.

Appendice A:
Descrizioni dei compiti degli attori AI dalle Figure 2 e 3

I compiti di AI Design sono svolti durante le fasi Application Context e Data and Input del ciclo di vita AI nella Figura 2. Gli attori AI Design creano il concetto e gli obiettivi dei sistemi AI e sono responsabili dei compiti di pianificazione, progettazione, raccolta e trattamento dati del sistema AI, affinché il sistema AI sia legittimo e idoneo allo scopo. I compiti includono articolare e documentare concetto e obiettivi del sistema, assunzioni sottostanti, contesto e requisiti; raccogliere e pulire dati; e documentare metadati e caratteristiche del dataset. Gli attori AI in questa categoria includono data scientist, esperti di dominio, analisti socio-culturali, esperti nel campo di diversità, equità, inclusione e accessibilità, membri di comunità impattate, esperti di fattori umani, ad esempio design UX/UI, esperti di governance, data engineer, fornitori di dati, finanziatori del sistema, product manager, entità terze, valutatori e governance legale e privacy.

I compiti di AI Development sono svolti durante la fase AI Model del ciclo di vita nella Figura 2. Gli attori AI Development forniscono l'infrastruttura iniziale dei sistemi AI e sono responsabili di compiti di costruzione e interpretazione del modello, che comportano creazione, selezione, calibrazione, training e/o testing di modelli o algoritmi. Gli attori AI in questa categoria includono esperti di machine learning, data scientist, sviluppatori, entità terze, esperti di governance legale e privacy ed esperti dei fattori socio-culturali e contestuali associati all'ambiente di distribuzione.

I compiti di AI Deployment sono svolti durante la fase Task and Output del ciclo di vita nella Figura 2. Gli attori AI Deployment sono responsabili delle decisioni contestuali relative a come il sistema AI è usato per assicurare la distribuzione del sistema in produzione. I compiti correlati includono pilotare il sistema, verificare compatibilità con sistemi legacy, assicurare conformità regolamentare, gestire il cambiamento organizzativo e valutare l'esperienza utente. Gli attori AI in questa categoria includono system integrator, sviluppatori software, utenti finali, operatori e practitioner, valutatori ed esperti di dominio con competenze in fattori umani, analisi socio-culturale e governance.

I compiti di Operation and Monitoring sono svolti nella fase Application Context/Operate and Monitor del ciclo di vita nella Figura 2. Questi compiti sono svolti da attori AI responsabili del funzionamento del sistema AI e della collaborazione con altri per valutare regolarmente output e impatti del sistema. Gli attori AI in questa categoria includono operatori di sistema, esperti di dominio, designer AI, utenti che interpretano o incorporano l'output dei sistemi AI, sviluppatori di prodotto, valutatori e auditor, esperti di compliance, management organizzativo e membri della comunità di ricerca.

I compiti di Test, Evaluation, Verification, and Validation (TEVV) sono svolti lungo l'intero ciclo di vita AI. Sono eseguiti da attori AI che esaminano il sistema AI o i suoi componenti, oppure rilevano e rimediano problemi. Idealmente, gli attori AI che svolgono compiti di verifica e validazione sono distinti da coloro che eseguono azioni di test e valutazione.

I compiti possono essere incorporati in una fase già dalla progettazione, dove i test sono pianificati in conformità ai requisiti di design.

- I compiti TEVV per design, pianificazione e dati possono concentrarsi su validazione interna ed esterna delle assunzioni per progettazione del sistema, raccolta dati e misurazioni rispetto al contesto previsto di distribuzione o applicazione.
- I compiti TEVV per sviluppo, cioè costruzione del modello, includono validazione e assessment del modello.
- I compiti TEVV per distribuzione includono validazione del sistema e integrazione in produzione, con testing e ricalibrazione per integrazione di sistemi e processi, esperienza utente e conformità a specifiche legali, regolamentari ed etiche esistenti.
- I compiti TEVV per operations implicano monitoraggio continuo per aggiornamenti periodici, testing e ricalibrazione dei modelli da parte di esperti di materia (SME), tracciamento degli incidenti o errori segnalati e loro gestione, rilevazione di proprietà emergenti e impatti correlati, e processi di ricorso e risposta.

I compiti e le attività di Human Factors si trovano lungo le dimensioni del ciclo di vita AI. Includono pratiche e metodologie di design human-centered, promozione del coinvolgimento attivo di utenti finali e altre parti interessate e attori AI rilevanti, incorporazione di norme e valori specifici del contesto nel design del sistema, valutazione e adattamento delle esperienze degli utenti finali e ampia integrazione di esseri umani e dinamiche umane in tutte le fasi del ciclo di vita AI. I professionisti dei fattori umani forniscono skill e prospettive multidisciplinari per comprendere il contesto d'uso, informare diversità interdisciplinare e demografica, partecipare a processi consultivi, progettare e valutare l'esperienza utente, svolgere valutazione e testing human-centered e informare gli impact assessment.

I compiti di Domain Expert comportano input da practitioner o studiosi multidisciplinari che forniscono conoscenza o competenza in, e su, un settore industriale, settore economico, contesto o area applicativa in cui un sistema AI è usato. Gli attori AI che sono esperti di dominio possono fornire guida essenziale per progettazione e sviluppo dei sistemi AI e interpretare output a supporto del lavoro svolto da team TEVV e di AI impact assessment.

I compiti di AI Impact Assessment includono assessment e valutazione dei requisiti per accountability del sistema AI, contrasto dei bias dannosi, esame degli impatti dei sistemi AI, sicurezza del prodotto, responsabilità legale e protezione, tra gli altri. Attori AI quali impact assessor e valutatori forniscono competenza tecnica, di fattori umani, socio-culturale e legale.

I compiti di Procurement sono condotti da attori AI con autorità finanziaria, legale o di policy management per l'acquisizione di modelli, prodotti o servizi AI da sviluppatori, vendor o contractor di terze parti.

I compiti di Governance and Oversight sono assunti da attori AI con autorità e responsabilità manageriale, fiduciaria e legale per l'organizzazione in cui un sistema AI è progettato, sviluppato e/o distribuito.

Gli attori AI chiave responsabili della governance AI includono management organizzativo, senior leadership e Board of Directors. Questi attori sono parti interessate all'impatto e alla sostenibilità dell'organizzazione nel suo complesso.

Attori AI aggiuntivi

Le entità terze includono provider, sviluppatori, vendor e valutatori di dati, algoritmi, modelli e/o sistemi e servizi correlati per un'altra organizzazione o per clienti dell'organizzazione. Le entità terze sono responsabili, in tutto o in parte, dei compiti di progettazione e sviluppo AI. Per definizione sono esterne al team di progettazione, sviluppo o distribuzione dell'organizzazione che acquisisce le loro tecnologie o servizi. Le tecnologie acquisite da entità terze possono essere complesse o opache e le tolleranze al rischio possono non allinearsi con quelle dell'organizzazione che distribuisce o gestisce.

Gli utenti finali di un sistema AI sono gli individui o gruppi che usano il sistema per scopi specifici. Questi individui o gruppi interagiscono con un sistema AI in un contesto specifico. Gli utenti finali possono variare per competenza da esperti AI a utenti finali di tecnologia alla prima esperienza.

Individui/comunità interessati comprendono tutti gli individui, gruppi, comunità o organizzazioni direttamente o indirettamente interessati da sistemi AI o da decisioni basate sull'output di sistemi AI. Tali individui non interagiscono necessariamente con il sistema o l'applicazione distribuita.

Altri attori AI possono fornire norme o linee guida formali o quasi formali per specificare e gestire i rischi AI. Possono includere associazioni di categoria, organizzazioni di sviluppo standard, gruppi di advocacy, ricercatori, gruppi ambientalisti e organizzazioni della società civile.

Il pubblico generale è il gruppo che con maggiore probabilità sperimenterà direttamente gli impatti positivi e negativi delle tecnologie AI. Può fornire la motivazione per le azioni intraprese dagli attori AI. Questo gruppo può includere individui, comunità e consumatori associati al contesto in cui un sistema AI è sviluppato o distribuito.

Appendice B:
Come i rischi AI differiscono dai rischi software tradizionali

Come per il software tradizionale, i rischi derivanti da tecnologie basate su AI possono essere più grandi di un'impresa, attraversare organizzazioni e produrre impatti sociali. I sistemi AI portano inoltre un insieme di rischi non affrontati in modo completo dagli attuali framework e approcci di rischio. Alcune caratteristiche dei sistemi AI che presentano rischi possono anche essere benefiche. Ad esempio, modelli pre-addestrati e transfer learning possono far avanzare la ricerca e aumentare accuratezza e resilienza rispetto ad altri modelli e approcci. Identificare fattori contestuali nella funzione MAP assisterà gli attori AI nel determinare il livello di rischio e i potenziali sforzi di gestione.

Rispetto al software tradizionale, i rischi specifici dell'AI nuovi o aumentati includono quanto segue:

- I dati usati per costruire un sistema AI possono non essere una rappresentazione vera o appropriata del contesto o dell'uso previsto del sistema AI, e la ground truth può non esistere o non essere disponibile. Inoltre, bias dannosi e altri problemi di qualità dei dati possono influenzare l'affidabilità del sistema AI, con possibili impatti negativi.
- Dipendenza e affidamento del sistema AI sui dati per i compiti di training, combinati con volume e complessità aumentati tipicamente associati a tali dati.
- Cambiamenti intenzionali o non intenzionali durante il training possono alterare fondamentalmente le prestazioni del sistema AI.
- I dataset usati per addestrare sistemi AI possono staccarsi dal loro contesto originario e previsto o diventare obsoleti rispetto al contesto di distribuzione.
- Scala e complessità dei sistemi AI, molti dei quali contengono miliardi o persino trilioni di punti decisionali, all'interno di applicazioni software più tradizionali.
- L'uso di modelli pre-addestrati, che può far avanzare la ricerca e migliorare le prestazioni, può anche aumentare i livelli di incertezza statistica e causare problemi di gestione del bias, validità scientifica e riproducibilità.
- Maggiore difficoltà nel prevedere modalità di guasto per proprietà emergenti di modelli pre-addestrati su larga scala.
- Rischio privacy dovuto alla capacità rafforzata di aggregazione dei dati dei sistemi AI.
- I sistemi AI possono richiedere manutenzione più frequente e trigger per condurre manutenzione correttiva a causa di drift dei dati, del modello o del concetto.
- Opacità aumentata e preoccupazioni sulla riproducibilità.

- Standard di testing software sottosviluppati e incapacità di documentare pratiche basate su AI secondo lo standard atteso dal software ingegnerizzato tradizionalmente, salvo i casi più semplici.
- Difficoltà nello svolgere testing regolare del software basato su AI, o nel determinare cosa testare, poiché i sistemi AI non sono soggetti agli stessi controlli dello sviluppo di codice tradizionale.
- Costi computazionali per sviluppare sistemi AI e loro impatto su ambiente e pianeta.
- Incapacità di prevedere o rilevare gli effetti collaterali dei sistemi basati su AI oltre le misure statistiche.

Considerazioni e approcci di gestione del rischio privacy e cybersecurity sono applicabili alla progettazione, sviluppo, distribuzione, valutazione e uso dei sistemi AI. I rischi privacy e cybersecurity sono considerati anche come parte di più ampie considerazioni di enterprise risk management, che possono incorporare rischi AI. Nell'ambito dello sforzo per affrontare caratteristiche di affidabilità AI quali "Secure and Resilient" e "Privacy-Enhanced", le organizzazioni possono considerare di sfruttare standard e linee guida disponibili che forniscono indicazioni ampie per ridurre rischi di sicurezza e privacy, inclusi, ma non limitati a, NIST Cybersecurity Framework, NIST Privacy Framework, NIST Risk Management Framework e Secure Software Development Framework. Questi framework hanno alcune caratteristiche in comune con l'AI RMF. Come la maggior parte degli approcci di gestione del rischio, sono outcome-based anziché prescrittivi e sono spesso strutturati attorno a un insieme Core di funzioni, categorie e sottocategorie. Sebbene vi siano differenze significative tra questi framework in base al dominio affrontato, e perché la gestione del rischio AI richiede di affrontare molti altri tipi di rischi, framework come quelli menzionati sopra possono informare considerazioni di sicurezza e privacy nelle funzioni MAP, MEASURE e MANAGE dell'AI RMF.

Allo stesso tempo, le linee guida disponibili prima della pubblicazione di questo AI RMF non affrontano in modo completo molti rischi dei sistemi AI. Ad esempio, framework e guide esistenti non sono in grado di:

- gestire adeguatamente il problema dei bias dannosi nei sistemi AI;
- affrontare i rischi sfidanti correlati all'AI generativa;
- affrontare in modo completo preoccupazioni di sicurezza relative a evasion, model extraction, membership inference, disponibilità o altri attacchi di machine learning;
- tenere conto della superficie d'attacco complessa dei sistemi AI o di altri abusi di sicurezza abilitati dai sistemi AI; e
- considerare rischi associati a tecnologie AI di terze parti, transfer learning e uso off-label, dove i sistemi AI possono essere addestrati per decision making fuori dai controlli di sicurezza di un'organizzazione o addestrati in un dominio e poi "fine-tuned" per un altro.

Sia l'AI sia le tecnologie e i sistemi software tradizionali sono soggetti a rapida innovazione. I progressi tecnologici dovrebbero essere monitorati e distribuiti per trarre vantaggio da tali sviluppi e lavorare verso un futuro dell'AI che sia insieme affidabile e responsabile.

Appendice C:
Gestione del rischio AI e interazione umano-AI

Le organizzazioni che progettano, sviluppano o distribuiscono sistemi AI per l'uso in ambienti operativi possono rafforzare la propria gestione del rischio AI comprendendo le limitazioni attuali dell'interazione umano-AI. L'AI RMF offre opportunità per definire e differenziare chiaramente i vari ruoli e responsabilità umani quando si usano, si interagisce con o si gestiscono sistemi AI.

Molti degli approcci data-driven su cui si basano i sistemi AI tentano di convertire o rappresentare pratiche osservative e decisionali individuali e sociali in quantità misurabili. Rappresentare fenomeni umani complessi con modelli matematici può comportare la rimozione del contesto necessario. Questa perdita di contesto può a sua volta rendere difficile comprendere impatti individuali e sociali che sono centrali per gli sforzi di gestione del rischio AI.

Questioni che meritano ulteriore considerazione e ricerca includono:

1. **Ruoli e responsabilità umani nel decision making e nella supervisione dei sistemi AI devono essere chiaramente definiti e differenziati.** Le configurazioni umano-AI possono variare da completamente autonome a completamente manuali. I sistemi AI possono prendere decisioni autonomamente, rinviare il decision making a un esperto umano o essere usati da un decisore umano come opinione aggiuntiva. Alcuni sistemi AI possono non richiedere supervisione umana, come modelli usati per migliorare la compressione video. Altri sistemi possono richiedere specificamente supervisione umana.
2. **Le decisioni che entrano nella progettazione, sviluppo, distribuzione, valutazione e uso dei sistemi AI riflettono bias sistemici e cognitivi umani.** Gli attori AI portano nel processo i propri bias cognitivi, individuali e di gruppo. I bias possono derivare dai compiti decisionali degli utenti finali ed essere introdotti lungo il ciclo di vita AI tramite assunzioni, aspettative e decisioni umane durante compiti di design e modellazione. Questi bias, che non sono necessariamente sempre dannosi, possono essere aggravati dall'opacità del sistema AI e dalla conseguente mancanza di trasparenza. Bias sistemici a livello organizzativo possono influenzare come i team sono strutturati e chi controlla i processi decisionali lungo il ciclo di vita AI. Questi bias possono anche influenzare decisioni a valle di utenti finali, decisori e policy maker e portare a impatti negativi.
3. **I risultati dell'interazione umano-AI variano.** In certe condizioni, ad esempio in compiti di giudizio basati sulla percezione, la componente AI dell'interazione umano-AI può amplificare bias umani, portando a decisioni più distorte rispetto all'AI o all'umano da soli.

3. Quando tali variazioni sono tuttavia considerate con giudizio nell'organizzazione di team umano-AI, possono produrre complementarità e miglioramento delle prestazioni complessive.
4. **Presentare informazioni del sistema AI agli esseri umani è complesso.** Gli esseri umani percepiscono e ricavano significato dall'output e dalle spiegazioni del sistema AI in modi diversi, riflettendo preferenze, tratti e competenze individuali differenti.

La funzione GOVERN offre alle organizzazioni l'opportunità di chiarire e definire ruoli e responsabilità degli esseri umani nelle configurazioni dei team umano-AI e di coloro che supervisionano le prestazioni del sistema AI. La funzione GOVERN crea inoltre meccanismi affinché le organizzazioni rendano più espliciti i propri processi decisionali, contribuendo a contrastare bias sistemici.

La funzione MAP suggerisce opportunità per definire e documentare processi relativi alla competenza di operatori e practitioner rispetto alle prestazioni del sistema AI e ai concetti di affidabilità, e per definire standard tecnici e certificazioni rilevanti. Implementare categorie e sottocategorie della funzione MAP può aiutare le organizzazioni a migliorare la propria competenza interna nell'analizzare il contesto, identificare limitazioni procedurali e di sistema, esplorare ed esaminare gli impatti dei sistemi basati su AI nel mondo reale e valutare i processi decisionali lungo il ciclo di vita AI.

Le funzioni GOVERN e MAP descrivono l'importanza dell'interdisciplinarità e di team demograficamente diversificati e dell'utilizzo di feedback da individui e comunità potenzialmente impattati. Gli attori AI richiamati nell'AI RMF che svolgono compiti e attività di fattori umani possono assistere i team tecnici ancorando pratiche di progettazione e sviluppo alle intenzioni degli utenti, a rappresentanti della più ampia comunità AI e ai valori sociali. Questi attori aiutano inoltre a incorporare norme e valori specifici del contesto nella progettazione del sistema e a valutare le esperienze degli utenti finali in relazione ai sistemi AI.

Gli approcci di gestione del rischio AI per configurazioni umano-AI saranno rafforzati da ricerca e valutazione continue. Ad esempio, il grado in cui gli esseri umani sono abilitati e incentivati a contestare l'output del sistema AI richiede ulteriori studi. Dati sulla frequenza e sulla razionale con cui gli esseri umani annullano l'output del sistema AI nei sistemi distribuiti possono essere utili da raccogliere e analizzare.

Appendice D:
Attributi dell'AI RMF

NIST ha descritto diversi attributi chiave dell'AI RMF quando il lavoro sul Framework è iniziato. Questi attributi sono rimasti intatti e sono stati usati per guidare lo sviluppo dell'AI RMF. Sono forniti qui come riferimento.

L'AI RMF mira a:

1. Essere basato sul rischio, efficiente nell'uso delle risorse, favorevole all'innovazione e volontario.
2. Essere guidato dal consenso e sviluppato e aggiornato regolarmente tramite un processo aperto e trasparente. Tutte le parti interessate dovrebbero avere l'opportunità di contribuire allo sviluppo dell'AI RMF.
3. Usare un linguaggio chiaro e semplice, comprensibile da un pubblico ampio, inclusi executive senior, funzionari governativi, leadership di organizzazioni non governative e persone che non sono professionisti AI, pur mantenendo profondità tecnica sufficiente per essere utile ai practitioner. L'AI RMF dovrebbe consentire la comunicazione dei rischi AI all'interno di un'organizzazione, tra organizzazioni, con clienti e al pubblico in generale.
4. Fornire linguaggio e comprensione comuni per gestire i rischi AI. L'AI RMF dovrebbe offrire tassonomia, terminologia, definizioni, metriche e caratterizzazioni per il rischio AI.
5. Essere facilmente utilizzabile e integrarsi bene con altri aspetti della gestione del rischio. L'uso del Framework dovrebbe essere intuitivo e prontamente adattabile come parte della più ampia strategia e dei processi di gestione del rischio di un'organizzazione. Dovrebbe essere coerente o allineato con altri approcci alla gestione dei rischi AI.
6. Essere utile a una vasta gamma di prospettive, settori e domini tecnologici. L'AI RMF dovrebbe essere universalmente applicabile a qualsiasi tecnologia AI e a casi d'uso specifici del contesto.
7. Essere focalizzato sui risultati e non prescrittivo. Il Framework dovrebbe fornire un catalogo di risultati e approcci anziché prescrivere requisiti one-size-fits-all.
8. Sfruttare e promuovere una maggiore consapevolezza di standard, linee guida, best practice, metodologie e strumenti esistenti per gestire i rischi AI, nonché illustrare la necessità di risorse aggiuntive e migliorate.
9. Essere agnostico rispetto a leggi e regolamenti. Il Framework dovrebbe supportare la capacità delle organizzazioni di operare sotto regimi legali o regolamentari nazionali e internazionali applicabili.
10. Essere un documento vivo. L'AI RMF dovrebbe essere prontamente aggiornato mentre tecnologia, comprensione e approcci all'affidabilità dell'AI e agli usi dell'AI cambiano e mentre le parti interessate apprendono dall'implementazione della gestione del rischio AI in generale e di questo framework in particolare.

Questa pubblicazione è disponibile gratuitamente all'indirizzo:
https://doi.org/10.6028/NIST.AI.100-1