FEATURE: ENGINEERING AI SYSTEMS RESPONSIBLY

RESPONSIBLE- AI-BY-DESIGN

A PATTERN COLLECTION FOR DESIGNING RESPONSIBLE ARTIFICIAL INTELLIGENCE SYSTEMS

GIUGNO 2024

FILE NAME: *IEEE - RESPONSIBLE AI BY* DESIGN - IT.

INDICE DEGLI ARGOMENTI

Titolo Pag.

I problemi di **IA RESPONSABILE** *si verificano spesso a livello di sistema, trasversale a molti componenti del sistema e all'intero ciclo di vita dell'ingegneria del software.*

Riassumiamo i modelli di progettazione che possono essere incorporati nei sistemi di IA come caratteristiche del prodotto per contribuire fin dalla progettazione.

Sebbene l'IA abbia un potenziale e una capacità significativa di stimolare la crescita economica e migliorare la produttività in una gamma crescente di settori, vi sono serie preoccupazioni sulla capacità dei sistemi di IA di comportarsi e prendere decisioni in modo responsabile.

Secondo un recente report di Gartner, il 21% delle organizzazioni ha già implementato o prevede di implementare tecnologie di IA responsabili entro i prossimi 12 mesi.

INTRODUCTION

Molti principi etici e linee guida sono stati recentemente emessi da governi, istituti di ricerca e aziende.

Tuttavia, questi principi sono di alto livello e difficilmente possono essere utilizzati nella pratica dagli sviluppatori.

La ricerca sull'**IA RESPONSABILE** si è concentrata su soluzioni algoritmiche limitate a un sottoinsieme di questioni, come l'EQUITÀ.

Le questioni etiche possono entrare in qualsiasi punto del ciclo di vita dell'ingegneria del software e sono spesso a livello di sistema, trasversale a molti componenti dei sistemi.

Per cercare di colmare il divario PRINCIPIO/ALGORITMO, sono iniziate ad apparire alcune linee guida di sviluppo.

Tuttavia, tali sforzi tendono a consistere in liste di controllo dei processi di sviluppo di alto livello e insiemi ad hoc privi di collegamenti relativi allo stato per i prodotti finali.

Pertanto, in questo articolo, piuttosto che rimanere a livello di PRINCIPIO ETICO o di ALGORITMO di IA, adottiamo un approccio orientato ai MODELLI e ci concentriamo sui MODELLI di PROGETTAZIONE a livello di sistema per integrare un'**IA RESPONSABILE** fin dalla progettazione nei prodotti finali.

I MODELLI DI PROGETTAZIONE sono raccolti sulla base dei risultati di una REVISIONE SISTEMATICA DELLA LETTERATURA (SYSTEMATIC LITERATURE REVIEW - SLR) e possono essere incorporati nella progettazione di sistemi di IA come caratteristiche del prodotto per contribuire a un'**IA RESPONSABILE** fin dalla progettazione.

Identifichiamo il CICLO DI VITA di un sistema di IA con provisioning in cui gli stati o le transizioni di stato sono associati ai MODELLI DI PROGETTAZIONE per mostrare quando tali modelli possono avere effetto.

Il CICLO DI VITA, insieme alle annotazioni dei MODELLI DI PROGETTAZIONE, fornisce una visione delle interazioni di sistema incentrata sull'**IA RESPONSABILE** e una guida per l'uso dei MODELLI DI PROGETTAZIONE per implementare l'**IA RESPONSABILE** dal punto di vista del sistema.

Per quanto ne sappiamo, questo è il primo studio che fornisce una guida concreta e attuabile alla progettazione a livello di sistema a cui architetti e sviluppatori possono fare riferimento.

METHODOLOGY

Per rendere operativa l'**IA RESPONSABILE**, abbiamo eseguito una SLR per identificare i MODELLI DI PROGETTAZIONE da utilizzare durante il PROCESSO DI SVILUPPO.

La Figura 1 illustra la metodologia.

La domanda della ricerca è la seguente: "Quali soluzioni per un'**IA RESPONSABILE** possono essere identificate?".

La domanda di ricerca si concentra sull'identificazione dei modelli riutilizzabili per un'**IA RESPONSABILE**.

Abbiamo usato "**IA**", "**RESPONSABILE**" e "**SOLUZIONE**" come termini chiave e abbiamo incluso sinonimi e abbreviazioni come termini supplementari per aumentare i risultati di ricerca.

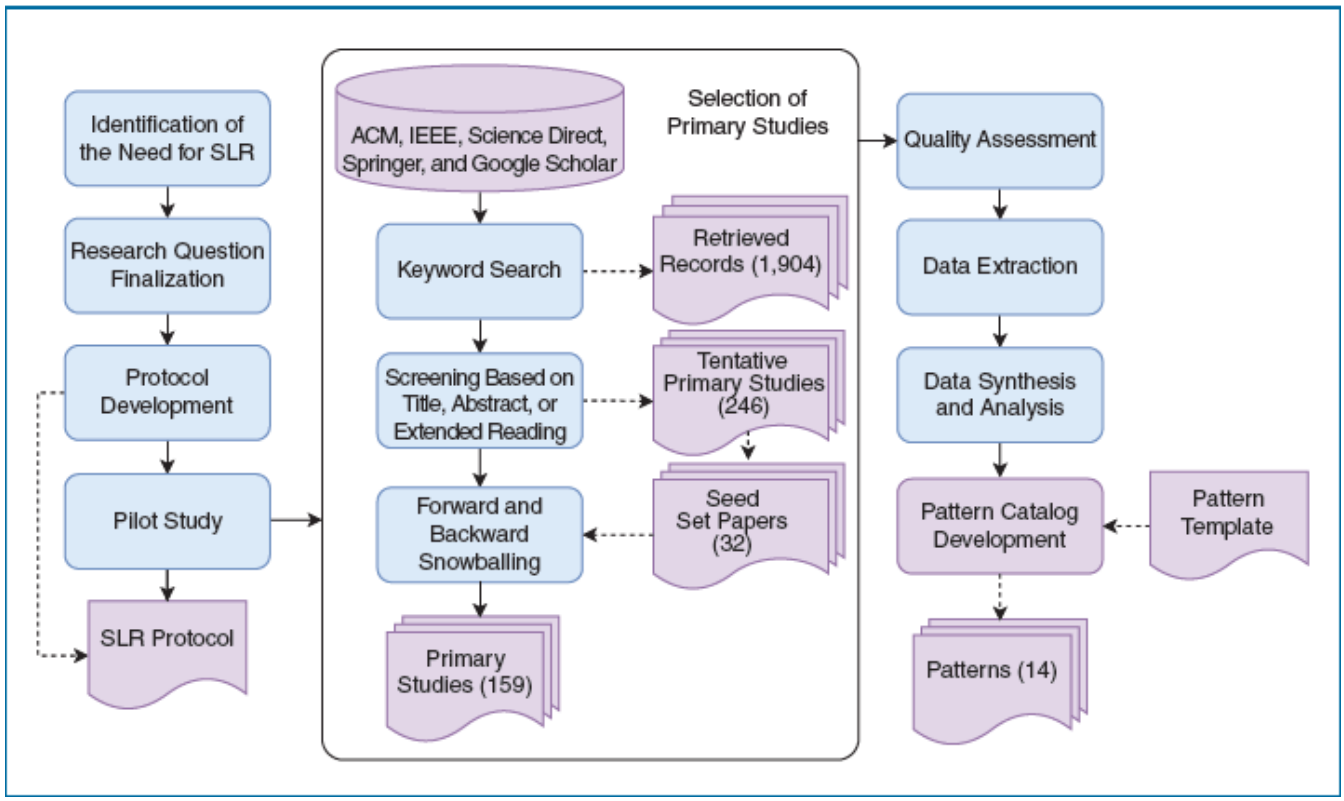
Le principali fonti di dati sono: il ASSOCIATION FOR COMPUTING MACHINERY DIGITAL LIBRARY, IEEE *XPLORE*, SCIENCE DIRECT, SPRINGER LINK, e GOOGLE SCHOLAR.

Lo studio include solo documenti e articoli che presentano soluzioni concrete di progettazione o di processo per un'**IA RESPONSABILE** ed esclude documenti e articoli che discutono solo di framework di alto livello.

È stata identificata una serie di 159 studi primari.

Utilizziamo i **PRINCIPI ETICI** elencati nello studio di mappatura dell'UNIVERSITÀ DI HARVARD: PRIVACY, RESPONSABILITÀ (la responsabilità professionale è fusa con responsabilità a causa delle definizioni sovrapposte), SICUREZZA e PROTEZIONE, TRASPARENZA e SPIEGABILITÀ, EQUITÀ e NON DISCRIMINAZIONE, CONTROLLO UMANO DELLA TECNOLOGIA e PROMOZIONE DEI VALORI UMANI.

FIG. 1 – THE METHODOLOGY – ACM: ASSOCIATION FOR COMPUTING MACHINERY



LIFECYCLE OF A PROVISIONED AI SYSTEM

La Figura 2 illustra il CICLO DI VITA DI UN SISTEMA DI IA di cui è stato eseguito il RIFORNIMENTO (PROVISIONING) utilizzando un diagramma di stato ed evidenzia i modelli associati a stati o transizioni pertinenti, che mostrano quando i MODELLI DI PROGETTAZIONE potrebbero avere effetto.

Abbiamo limitato l'ambito ai MODELLI DI PROGETTAZIONE che possono essere incorporati nei sistemi di IA e alla catena di strumenti della CATENA DI APPROVVIGIONAMENTO (SUPPLY CHAIN) fornita come caratteristiche del prodotto finale.

Le procedure consigliate del PROCESSO DI SVILUPPO, inclusi alcuni modelli correlati al **training** del modello offline, non rientrano nell'ambito di questo articolo.

Prima del PROVISIONING di un sistema di IA, è possibile accedere alle informazioni sulla SUPPLY CHAIN tramite la DISTINTA BASE (BILL OF MATERIALS).

Agli utenti può essere richiesto di fornire le **CREDENZIALI ETICHE VERIFICABILI** per dimostrare la loro capacità di utilizzare i sistemi e gli utenti possono esaminare le **CREDENZIALI ETICHE VERIFICABILI** del sistema per il controllo della **CONFORMITÀ ETICA**.

Una volta che il sistema di IA inizia a funzionare, è importante eseguire la simulazione a livello di sistema attraverso un **DIGITAL TWIN ETICO**.

Una SANDBOX etica può essere utilizzata per separare fisicamente i componenti di IA da quelli non di IA.

Quando a un sistema di IA è richiesto di eseguire un'attività, spesso è necessario prendere decisioni prima di eseguirla.

Un componente IA può essere attivato o disattivato tramite un commutatore di modelli AI per prendere automaticamente la decisione o coinvolgere esperti umani per rivedere il suggerimento.

Un decisore multi-modello può utilizzare modelli diversi per prendere una singola decisione e controllare i risultati.

Allo stesso modo, la ridondanza omogenea può essere applicata alla progettazione del sistema per consentire la tolleranza ai guasti.

Sia i comportamenti sia i risultati decisionali del sistema di IA sono monitorati e validati attraverso un VALIDATORE ETICO CONTINUO.

Gli incentivi per i comportamenti etici possono essere mantenuti da un REGISTRO DEGLI INCENTIVI.

Se il sistema non soddisfa i requisiti (compresi i requisiti etici) o viene rilevato un quasi incidente, il sistema deve essere aggiornato.

Un tecnico ripete il training del modello localmente in ogni client per proteggere la privacy dei dati.

Il REGISTRO DI CONTROLLO DELLE VERSIONI può essere utilizzato per tenere traccia della coevoluzione dei componenti o delle risorse del sistema di IA.

È possibile creare una BASE DI CONOSCENZE ETICHE per rendere sistematicamente accessibili e utilizzate tali conoscenze durante lo sviluppo o l'aggiornamento del sistema.

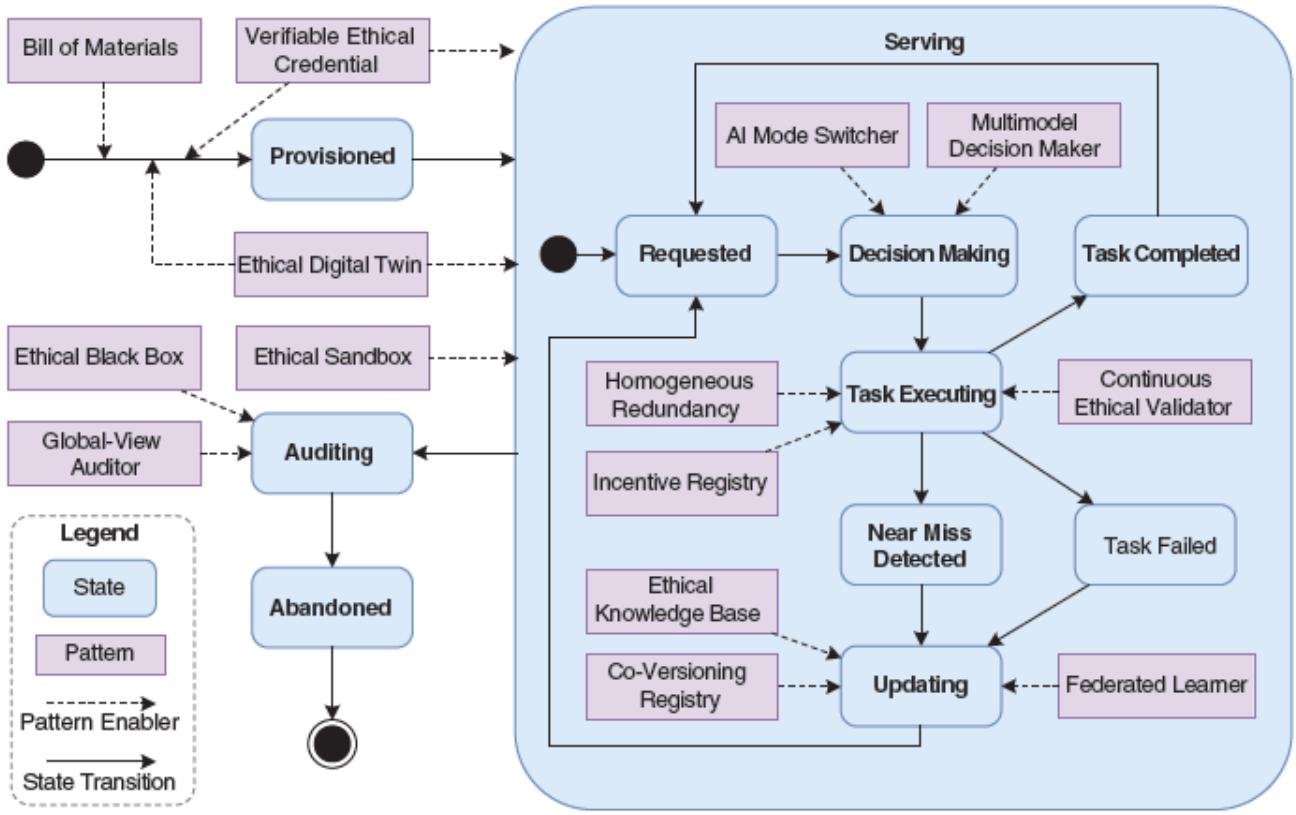
Il sistema deve essere sottoposto a controlli periodici o quando si verificano gravi guasti/quasi incidenti.

Una SCATOLA NERA ETICA può essere progettata per registrare i dati critici che possono essere conservati come prova.

Un "**revisore generale**" **(global-view auditor)** può essere costruito per fornire una visione olistica della RESPONSABILITÀ quando più sistemi sono coinvolti in un incidente

Le parti interessate possono decidere se abbandonare il sistema di IA se non soddisfa più i requisiti.

FIG. 2 – THE LIFECYCLE OF A PROVISIONED AI SYSTEM



DESIGN PATTERN (SCHEMA DI PROGETTAZIONE)

Per rendere operativa l'**IA RESPONSABILE**, la Figura 3 elenca una raccolta di modelli per l'**IA RESPONSABILE** FIN DALLA PROGETTAZIONE (RESPONSIBLE-AI-BY-DESIGN).

FIG. 3 – OPERATIONALIZED DESIGN PATTERNS FOR RESPONSIBLE AI SYSTEMS

(Account: accountability; Explain.: explainability; N/A: not applicable; Transp.: transparency)

	Context	Problem	Degree of relevance to principles							Solution	Related patterns	Consequences	
Pattern name	Impacted stakeholders	Type of objective	Privacy	Account.	Safety and security	Traspa. and explain.	Fairness	Human control	Human values	Mechanism name	Pattern name	Benefits	Drawbacks
Bill of materials	Development teams, RAI governors, AI users, AI consumers	Trust	N/A	Yes	Yes	Yes	N/A	N/A	N/A	Immutable data infrastructure, context documents	Verifiable ethical credential	Increased transparency, increased accountability, integrity	Increased management effort
Verifiable ethical credentials	Development teams, RAI governors, AI users, AI consumers	Trust	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Publicly accessbile data infrastructure	Bill of materials	Increased trust, AI system adoption, awareness of AI ethical issues	Set-once-and-forget, human-in the loop, interoperability, authenticity
Ethical digital twin	Operators, data scientists	Trustworthiness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Simulation infrastructure	Ethical sandbox, AI mode switcher, ethical knowledge base	Cost-efficiency, increased ethical quality of outcomes of AI systems	Limited by quality of the simulation model, increased cost
Ethical sandbox	Data scientists	Trustworthiness	Yes	N/A	Yes	N/A	Yes	N/A	Yes	Ethical margin, watchdog	Ethical digital twin, AI mode switcher	Increased ethical quality, safety	Applicability, performance penalty
AI mode switcher	Operators, AI users, AI consumers	Trustworthiness, trust	N/A	N/A	Yes	N/A	N/A	Yes	Yes	Invocation, dismissal, kill switch, override, fallback, built-in guard, recourse channel	Ethical digital twin, ethical sandbox	Increased trust, contestability and autonomy	Efficiency, suitability to (near) real time systems
Multimodel decision-maker	AI users, AI consumers	Trustworthiness	N/A	N/A	Yes	N/A	Yes	N/A	N/A	Unified interface	Homogeneous redundancy	Increased reliability, fairness	Increased development effort, required more skills, decreased training efficiency
Homogeneous redundancy	Data scientists	Trustworthiness	N/A	N/A	Yes	N/A	N/A	Yes	N/A	Unified interface	Multimodel decision maker	Fault tolerance, increased safety and human control	Increased operating cost, performance penalty
Incentive registry	Data scientists	Trustworthiness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Publicly accessbile data infrastructure	Continuous ethical validator, federated learner	Increased motivation for ethical behavior or decisions	Limitation of the incentive design, potential privacy breach risk
Continuous ethical validator	Operators, data scientists	Trustworthiness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Version-based feedback, rebuild alert	Incentive registry, ethical knowledge base	Increased maintainability	Suitability to all ethical risks
Ethical knowledge base	RAI governors, AI users, AI consumers	Trustworthiness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Knowledge graph	Ethical digital twin, continuous ethical validator	Compliance checking	Increased development effort
Coversioning registry	Operators, developers, data scientists	Trustworthiness, trust	N/A	Yes	N/A	Yes	N/A	N/A	N/A	Immutable data infrastructure	Federated learner	Traceability and accountability	Complexity
Federated learner	Developers, operators	Trustworthiness	Yes	N/A	Yes	N/A	N/A	N/A	N/A	Asynchronous/hierarchical/decentralised/secure aggregator	Incentive registry, coversioning registry, global-view auditor	Privacy, increased reliability	Sampling bias, performance penalty
Ethical black box	RAI governors, AI users, AI consumers	Trust	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Immutable log	Global-view auditor	Accountability, traceability	Privacy
Global-view auditor	RAI governors, AI users, AI consumers	Trust	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Immutable log	Federated learner, ethical black box	Accountability, traceability	Performance

BILL OF MATERIALS (DISTINTA BASE)

I fornitori di prodotti di IA spesso creano sistemi assemblando componenti commerciali o open source e/o non di IA di terze parti.

Gli utenti hanno spesso **PREOCCUPAZIONI ETICHE** riguardanti i sistemi/componenti di IA ottenuti.

Prima del PROVISIONING di un sistema, è possibile accedere alle informazioni sulla SUPPLY CHAIN tramite la DISTINTA BASE, che mantiene una registrazione formale, e leggibile dalla macchina, dei dettagli della CATENA DI APPROVVIGIONAMENTO dei componenti utilizzati nella creazione di un sistema di IA, come: il nome del componente, la versione, il fornitore, la relazione di dipendenza, l'autore e il timestamp.

Scopo della DISTINTA BASE è quello di fornire tracciabilità e trasparenza nei componenti che compongono i sistemi affinché le questioni etiche possano essere tracciate e affrontate.

Ci sono stati molti strumenti per generare una distinta base software per i professionisti, come DEPENDENCY-TRACK.

Per garantire la TRACCIABILITÀ e l'INTEGRITÀ, è necessaria un'infrastruttura di dati immutabile per archiviare i dati della distinta base.

Ad esempio, i produttori di veicoli autonomi possono mantenere un contratto di registro dei materiali su blockchain per tenere traccia delle informazioni sulla catena di approvvigionamento dei loro componenti, ad esempio la versione e il fornitore del componente di navigazione di terze parti basato sull'intelligenza artificiale.

VERIFIABLE ETHICAL CREDENTIALS

Le CREDENZIALI ETICHE VERIFICABILI sono dati verificabili crittograficamente che possono essere utilizzati come prova solida della conformità etica per i sistemi, dei componenti, degli artefatti e delle parti interessate (come sviluppatori e utenti).

Prima di utilizzare i sistemi, gli utenti devono verificare le CREDENZIALI ETICHE dei sistemi per accertare se tali sistemi sono conformi ai PRINCIPI o alle NORMATIVE ETICHE.

D'altra parte, agli utenti è spesso richiesto di fornire le CREDENZIALI ETICHE per utilizzare e far funzionare i sistemi.

È necessario creare un'infrastruttura di dati accessibile al pubblico per supportare la GENERAZIONE e la VERIFICA delle CREDENZIALI ETICHE.

Ad esempio, prima di guidare un veicolo, al conducente viene richiesto di scansionare le proprie credenziali etiche per dimostrare di avere la capacità di guidare in sicurezza, verificando al

contempo le credenziali etiche del sistema di guida automatizzata del veicolo mostrate sulla console centrale.

ETHICAL DIGITAL TWIN

Prima di eseguire il sistema di IA PROVISIONED in un ambiente di produzione, è fondamentale condurre una simulazione a livello di sistema attraverso un DIGITAL TWIN ETICO (ETHICAL DIGITAL TWIN) in esecuzione su una piattaforma di simulazione per monitorare i comportamenti dei sistemi e prevedere potenziali RISCHI ETICI.

Un ETHICAL DIGITAL TWIN può anche essere progettato come componente a livello di infrastruttura operativa per esaminare i comportamenti e le decisioni di runtime di un sistema sulla base del modello di simulazione astratto utilizzando dati del mondo reale.

I risultati della VALUTAZIONE DEL RISCHIO (RISK ASSESSMENT – RA) possono essere utilizzati dal sistema o dagli utenti per intraprendere ulteriori azioni volte a mitigare i potenziali RISCHI ETICI.

Ad esempio, i produttori di veicoli autonomi possono utilizzare il DIGITAL TWIN ETICO per esplorare i limiti dei veicoli autonomi sulla base dei dati di runtime raccolti, come NVIDIA DRIVE SIM e XFPRO.

ETHICAL SANDBOX

È RISCHIOSO ESEGUIRE L'INTERO SISTEMA, COMPRESI I COMPONENTI AI E I COMPONENTI NON AI, NELLO STESSO AMBIENTE.

Quando è proposto un sistema, una SANDBOX ETICA può essere utilizzata per separare fisicamente i componenti di IA dai componenti non di IA eseguendo i componenti di IA in un ambiente di esecuzione di emulazione autonomo, ad esempio il SANDBOXING del componente di percezione visiva non verificato.

I componenti collocati nella SANDBOX ETICA non hanno accesso al resto del sistema.

Tutte le funzionalità hardware e software dei componenti sono duplicate nella SANDBOX ETICA.

Pertanto, i componenti possono essere eseguiti in modo sicuro sotto supervisione prima di essere distribuiti su larga scala.

Ad esempio, FASTCASE AI SANDBOX fornisce una piattaforma di esecuzione sicura per analizzare i dati in modo sicuro in un ambiente sicuro.

La probabilità massima tollerabile dovrebbe essere impostata come margine etico per la SANDBOX rispetto ai REQUISITI ETICI.

È possibile aggiungere un ***watchdog*** per limitare il tempo di esecuzione del componente per evitare il potenziale RISCHIO ETICO, ad esempio eseguendo il componente di percezione visiva solo per 10 minuti su strade progettate appositamente per i veicoli autonomi.

AI MODE SWITCHER

Un'importante decisione di PROGETTAZIONE ARCHITETTURALE, in fase di progettazione di un sistema, è il momento in cui si attiva il SISTEMA DI IA.

Quando un sistema prende una decisione, il selettore di modalità di IA consente meccanismi efficienti di chiamata e di disattivazione allo scopo di attivare o arrestare la componente IA quando necessario.

Il KILL SWITCH è un tipo speciale di meccanismo di chiamata che disattiva immediatamente la componente IA e termina i suoi effetti negativi, ad esempio disattivando la funzionalità dell'autopilota e la sua connessione Internet.

La componente IA può prendere decisioni automaticamente o fornire suggerimenti a esperti umani in situazioni ad alto rischio.

Le decisioni possono essere approvate o annullate da un esperto (ad esempio, saltando il percorso suggerito dal sistema di navigazione).

Se lo stato del sistema dopo aver agito su una decisione non è previsto dagli esperti, è possibile attivare un FALL-BACK per riportare il sistema allo stato precedente.

Una protezione integrata garantisce che il componente sia utilizzato solo nelle categorie di rischio predefinite.

MULTI-MODEL DECISION MAKER

L'AFFIDABILITÀ del software tradizionale dipende dalla progettazione dei componenti software.

Una delle pratiche di AFFIDABILITÀ è la RIDONDANZA, che può essere applicata ai componenti di IA.

Quando vengono prese decisioni da un sistema, un decisore multi-modello può eseguire diversi modelli per prendere un'unica decisione, ad esempio utilizzando algoritmi diversi per la percezione visiva.

L'AFFIDABILITÀ può essere migliorata utilizzando modelli diversi in contesti diversi (ad esempio, gruppi di utenti o regioni diverse).

Inoltre, la tolleranza di errore può essere abilitata effettuando un controllo incrociato dei risultati forniti da più modelli (ad esempio, accettando solo gli stessi risultati dai modelli distribuiti).

IBM WATSON NATURAL LANGUAGE UNDERSTANDING effettua previsioni utilizzando un framework di apprendimento d'insieme che include più modelli di rilevamento delle emozioni.

HOMOGENEOUS REDUNDANCY

I fallimenti etici nei sistemi possono causare gravi danni all'uomo o all'ambiente.

La programmazione della N-VERSION è un modello di progettazione per affrontare i problemi di AFFIDABILITÀ del software tradizionale.

Questo concetto può essere adattato e applicato alla progettazione di sistemi di IA.

La RIDONDANZA OMOGENEA (ad esempio, due componenti di controllo dei freni) può essere applicata per tollerare i componenti del sistema di IA altamente incerti che possono prendere decisioni non etiche o i componenti hardware avversari che producono dati dannosi o si comportano in modo non etico.

Quando un sistema di IA esegue un'attività, è possibile eseguire un controllo incrociato per gli output forniti da più componenti ridondanti di un unico tipo.

INCENTIVE REGISTRY

Gli incentivi sono efficaci per motivare i sistemi a svolgere compiti in modo **responsabile**.

Durante l'esecuzione di un'attività, un REGISTRO DEGLI INCENTIVI registra le RICOMPENSE (REWARDS) che vengono assegnate per le decisioni e i comportamenti dei sistemi, ad esempio le ricompense per il percorso consigliato senza rischi per la sicurezza.

Esistono diversi modi per applicare il MECCANISMO DI INCENTIVAZIONE, ad esempio progettandolo in un'infrastruttura di dati basata su blockchain accessibile pubblicamente utilizzando l'APPRENDIMENTO PER RINFORZO (REINFORCEMENT LEARNING).

Tuttavia, è difficile progettare i meccanismi in un contesto di **IA RESPONSABILE** poiché è difficile misurare l'IMPATTO ETICO delle decisioni e dei comportamenti dei sistemi su alcuni **principi etici (come i valori umani)**.

Inoltre, è necessario che tutte le parti interessate raggiungano un consenso sul MECCANISMO DI INCENTIVAZIONE.

In alcuni casi, i PRINCIPI ETICI sono in conflitto tra loro, rendendo più difficile la progettazione di un MECCANISMO DI INCENTIVAZIONE.

FLOBC è uno strumento che utilizza la blockchain per incentivare i contributi formativi per l'apprendimento.

CONTINUOUS ETHICAL VALIDATOR

I sistemi spesso devono condurre un apprendimento continuo quando si rilevano spostamenti di dati o comportamenti non etici in produzione.

Quando un sistema esegue i compiti, un VALIDATORE ETICO CONTINUO MONITORA e CONVALIDA i risultati dei sistemi (ad esempio, il percorso suggerito dal sistema di navigazione) rispetto ai REQUISITI ETICI.

I risultati dei sistemi sono le conseguenze delle decisioni e dei comportamenti dei sistemi stessi, ovvero se il sistema si comporta eticamente o fornisce i benefici promessi in una determinata situazione.

L'ora e la frequenza della convalida possono essere predefinite all'interno del VALIDATORE CONTINUO.

È possibile inviare un feedback basato sulla versione e un avviso di ricostruzione quando i requisiti etici vengono soddisfatti o violati.

Un REGISTRO DEGLI INCENTIVI può essere utilizzato per "PREMIARE O PUNIRE" i comportamenti o le decisioni etiche/non etiche dei sistemi.

ETHICAL KNOWLEDGE BASE

I sistemi di IA implicano un'ampia conoscenza etica, compresi i PRINCIPI ETICI DELL'IA, i regolamenti, i casi d'uso non etici, ecc.

Sfortunatamente, tale conoscenza etica è sparsa in diversi documenti (ad esempio, incidenti di IA) ed è di solito implicita o addirittura sconosciuta agli sviluppatori, che si concentrano principalmente sugli aspetti tecnici e non hanno un background etico.

Ciò si traduce in negligenza o nell'uso ad hoc di conoscenze etiche pertinenti nello sviluppo di sistemi.

Una base di conoscenza etica si basa su un grafo della conoscenza per rendere esplicite e tracciabili entità significative, concetti e le loro ricche relazioni semantiche attraverso documenti eterogenei, in modo che la conoscenza etica possa essere sistematicamente accessibile, analizzata e utilizzata durante lo sviluppo o l'aggiornamento dei sistemi.

Ad esempio, una base di conoscenze etiche può essere utilizzata per supportare una VALUTAZIONE CONTINUA DEL RISCHIO ETICO.

È possibile costruire una base di conoscenze etiche sulla base dei principi e dei quadri etici dell'IA, nonché dei casi d'uso effettivi dell'IA discussi nei documenti e negli articoli esistenti.

CO-VERSIONING REGISTRY

I sistemi di IA comportano diversi livelli di dipendenza e necessitano di un'evoluzione frequente quando si verifica la deriva dei dati o un comportamento non etico.

Il co-versioning dei componenti dei sistemi o delle risorse generate nelle pipeline fornisce garanzie di provenienza per l'intero ciclo di vita dei sistemi.

Sono stati utilizzati molti strumenti di controllo delle versioni per la gestione del co-versioning di dati e modelli, come DVC (DATA VERSION CONTROL).

Quando si aggiorna un sistema di IA, un REGISTRO DI CONTROLLO DELLE VERSIONI può tenere traccia della co-evoluzione dei componenti o delle risorse di IA.

Esistono diversi livelli di co-versioning:

- ✓ di componenti AI,
- ✓ di componenti non AI,
- ✓ delle risorse all'interno dei componenti AI (ad esempio, co-versioning di dati, modello, codice e configurazioni).

Un'infrastruttura di dati accessibile pubblicamente può essere utilizzata per gestire il REGISTRO DELLE VERSIONI condiviso per fornire una traccia attendibile per le dipendenze.

Ad esempio, un contratto di registro di co-versioning può essere creato su blockchain per gestire diverse versioni dei modelli di percezione visiva e i corrispondenti set di dati di addestramento.

FEDERATED LEARNER

Nonostante i dispositivi mobili o Internet of Things ampiamente diffusi generino enormi quantità di dati, la fame di dati è ancora una sfida, date le crescenti preoccupazioni relative alla privacy dei dati.

Durante l'apprendimento o l'aggiornamento dei modelli, un tecnico preserva la privacy dei dati eseguendo l'addestramento del modello localmente sui dispositivi client e formulando un modello globale su un server centrale basato sugli aggiornamenti del modello locale, ad esempio addestrando il modello di percezione visiva localmente in ciascun veicolo.

L'apprendimento decentralizzato è un'alternativa all'apprendimento federato che utilizza la blockchain per rimuovere il singolo punto di errore e coordinare il processo di apprendimento in modo completamente decentralizzato.

In caso di esiti negativi, gli esseri umani responsabili possono essere rintracciati e identificati da una scatola nera etica per la responsabilità.

ETHICAL BLACK BOX

La scatola nera è stata introdotta inizialmente per gli aerei diversi decenni fa per la registrazione dei dati di volo critici.

Scopo dell'incorporazione di una SCATOLA NERA ETICA in un sistema è quello di controllare il sistema e indagare perché e come il sistema ha causato un incidente o una possibilità che si verificasse.

La SCATOLA NERA ETICA registra continuamente i dati dei sensori, dello stato interno, delle decisioni, dei comportamenti (sia del sistema che dell'operatore) e degli effetti.

Ad esempio, una SCATOLA NERA ETICA potrebbe essere integrata nel sistema di guida automatizzata per registrare i comportamenti del sistema e del conducente e i loro effetti.

È necessario prendere decisioni di progettazione su quali dati devono essere registrati e dove devono essere archiviati (ad esempio, utilizzando un registro immutabile basato su blockchain o un'archiviazione dei dati basata su cloud).

GLOBAL-VIEW AUDITOR

Ci può essere più di un sistema di IA coinvolto in un INCIDENTE ETICO (ad esempio, più veicoli autonomi in un incidente d'auto).

Durante l'auditing, è spesso difficile identificare la responsabilità, poiché i dati raccolti da ciascuno dei sistemi coinvolti possono entrare in conflitto tra loro.

Un revisore con una visione globale può consentire la responsabilità analizzando le discrepanze dei dati tra i sistemi di IA coinvolti e identificando la responsabilità per l'incidente.

Questo modello può essere applicato anche per migliorare l'AFFIDABILITÀ di un sistema prendendo i dati da altri sistemi.

Ad esempio, un veicolo autonomo aumenta la sua visibilità utilizzando i dati di percezione raccolti da altri veicoli.

Tutti i dati storici dei sistemi possono essere inseriti in un registro per l'audit di terze parti.

CONCLUSION

Per rendere operativa l'**IA RESPONSABILE**, adottiamo un approccio orientato ai MODELLI e raccogliamo una serie di MODELLI di PROGETTAZIONE del prodotto incorporabili in un sistema come CARATTERISTICHE DEL PRODOTTO per consentire un'**IA RESPONSABILE** FIN DALLA PROGETTAZIONE (RESPONSIBLE AI BY DESIGN).

I MODELLI sono associati agli STATI o alle TRANSIZIONI di STATO di un sistema con PROVISIONING, fungendo da guida efficace per progettare il **SISTEMA RESPONSABILE**.

Attualmente stiamo costruendo un CATALOGO DI MODELLI DI IA (AI PATTERN CATALOG) RESPONSABILE che include modelli di:

1. GOVERNANCE multi-livello,
2. PROCESSO affidabile (ad esempio, best practice e tecniche),
3. PRODOTTO di **IA RESPONSABILE** FIN DALLA PROGETTAZIONE (RESPONSIBLE- AI-BY-DESIGN).