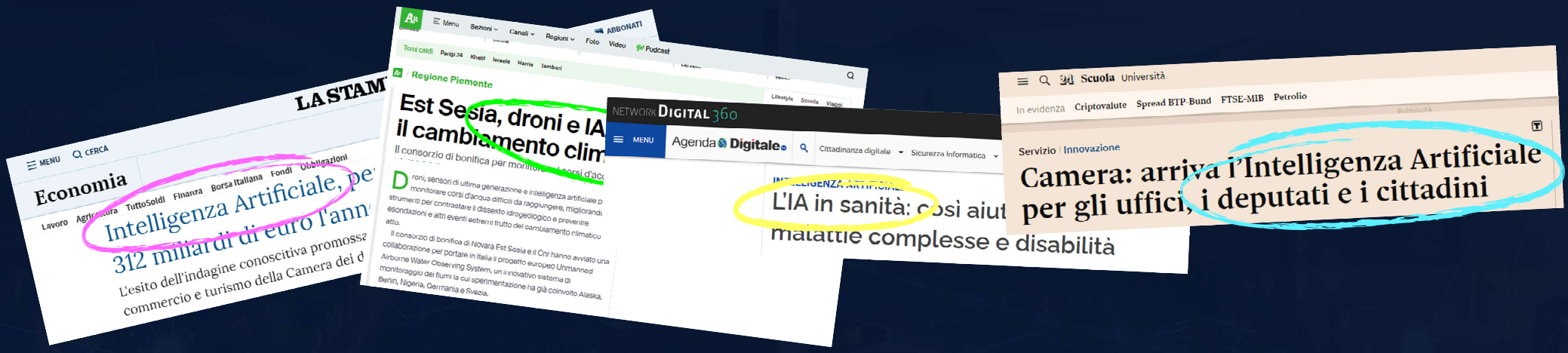
La Via Italiana dell'Intelligenza Artificiale

26 Marzo 2025

FASTWEB + vodafone

L’AI Generativa sta entrando sempre di più nelle nostre vite…



2

… ma l'utilizzo di questi strumenti porta un concreto rischio di perdita di controllo sui dati

Utilizzare gli strumenti di AI Generativa ad oggi liberamente accessibili (ChatGPT, ChatPDF, …) significa inviare **dati al di fuori dell'UE** e perderne il controllo

Perche’ parliamo di un LLM Nazionale?

FASTWEB

Il nostro Modello Linguistico Nazionale

FASTWEB MIA

FASTWEB

AI Security

Rischi di Sicurezza

La sicurezza è un aspetto critico a fronte di vulnerabilità come gli «adversary attack» che minacciano la reputazione, privacy dei dati e affidabilità dei sistemi aziendali. I principali **tipi di attacchi** per modelli di AI/LLM sono:

Impersonificazione e Attacchi di Social Engineering

I modelli possono essere usati per imitare persone reali e ingannare utenti (ad esempio: un deepfake testuale che finge di essere il CEO di un'azienda per frodare i dipendenti)

Disinformazione e Manipolazione

I modelli possono essere usati per generare fake news, deepfake testuali o contenuti manipolati (ad esempio: un modello usato per scrivere email di phishing personalizzate con dati reali)

Uso Malevolo da Parte di Criminali

I modelli possono essere utilizzati per scopi illeciti, come scrittura di malware o codice exploit, generazione di documenti falsificati o automazione di truffe o attacchi informatici (ad esempio: un LLM usato per scrivere codice di un ransomware sofisticato)

Adversary Attack

Gli aggressori manipolano gli input per far sì che i modelli producano risposte errate o pericolose (ad esempio: un prompt studiato per aggirare i filtri e generare contenuti vietati)

Bias e Discriminazione

I modelli se addestrati su dati non bilanciati, possono amplificare pregiudizi sociali, decisioni discriminatorie in AI o propagazione di stereotipi di genere o razziali (ad esempio: un chatbot AI che fornisce risposte discriminatorie su base etnica o di genere)

Esfiltrazione di Dati Sensibili

I modelli potrebbero memorizzare e rivelare informazioni riservate, come dati aziendali o personali (ad esempio: chiedere ripetutamente un numero di carta di credito nascosto nei dati di training)

Rischi di Sicurezza

La sicurezza è un aspetto critico a fronte di vulnerabilità come gli *adversary attack* che minacciano la reputazione, privacy dei dati e affidabilità dei sistemi aziendali. I principali **tipi di attacchi** per modelli di AI/LLM sono:

RISCHI INDIPENDENTI DAI MODELLI AI/LLM

Impersonificazione e Attacchi di Social Engineering

I modelli di GenAI facilitano questo tipo di attacchi, aumentando la credibilità degli attaccanti agli occhi delle vittime attraverso l'imitazione realistica di persone reali (ad esempio: un deepfake testuale che finge di essere il CEO di un'azienda per frodare i dipendenti)

Disinformazione e Manipolazione

I modelli possono essere usati per generare fake news, deepfake testuali o contenuti manipolati (ad esempio: una fake news come "WhatsApp diventa a pagamento" induce la vittima a cliccare su un link fraudolento e a effettuare un pagamento)

Uso Malevolo da Parte di Criminali

I modelli possono essere utilizzati per scopi illeciti, come scrittura di codice exploit o malware ex-novo (che rende la detection pre-esecuzione difficoltosa), generazione di documenti falsificati o automazione di truffe o attacchi informatici (ad esempio: un LLM usato per scrivere codice di un ransomware sofisticato)

RISCHI A CUI SONO ESPOSTI I MODELLI AI/LLM

Adversary Attack

Gli aggressori manipolano gli input per far sì che i modelli producano risposte errate o pericolose (ad esempio: un prompt studiato per aggirare i filtri e generare contenuti vietati)

Bias e Discriminazione

I modelli se addestrati su dati non bilanciati, possono amplificare pregiudizi sociali, decisioni discriminatorie in AI o propagazione di stereotipi di genere o razziali (ad esempio: un modello AI per il profiling di minacce terroristiche potrebbe escludere alcune persone a causa di bias nel training, valutando in modo errato alcuni dati)

Esfiltrazione di Dati Sensibili

I modelli potrebbero memorizzare e rivelare informazioni riservate, come dati aziendali o personali (ad esempio: chiedere ripetutamente un numero di carta di credito nascosto nei dati di training)

Qual è l‘approccio di Fastweb?

Fastweb ha sviluppato internamente **soluzioni AI/LLM adottando l’approccio *Secure by Design*** che in maniera strutturata copre ogni fase del ciclo di vita dei modelli: dallo sviluppo al deployment e al monitoraggio continuo.

Security by Design

Integrazione della sicurezza fin dalla **fase di progettazione** per prevenire vulnerabilità, ridurre i rischi di attacchi e garantire un uso responsabile del modello

Difesa contro Attacchi

Protezione dei modelli **da manipolazioni e attacchi** malevoli che tentano di aggirare le restrizioni:

- Mitigando i prompt injection
- Proteggendo contro i data poisoning
- Addestrando i modelli per renderli più resistente a input ingannevoli

Red Teaming

Esecuzione di simulazioni di attacchi reali per **testare** e mettere alla prova la **resistenza dei modelli** AI/LLM contro minacce e vulnerabilità che possano essere sfruttate da attaccanti.
Questa pratica combina i tradizionali *adversial testing* con metodologie specifiche per l’AI, affrontando rischi come la prompt injection, gli output tossici, l'estrazione di modelli, le distorsioni, rischi di conoscenza e allucinazioni.

Privacy e Protezione dei Dati

Evitare che il modello divulghi **informazioni sensibili** o venga usato in modo improprio con:

- Filtri di output per impedire la generazione di dati privati
- Access Control e API Security per limitare l’uso non autorizzato del modello
- Tracciabilità e watermarking per identificare i contenuti generati dall’AI

Monitoraggio e Aggiornamenti Continui

Implementazione di un **monitoraggio continuo** e **aggiornamenti regolari** tramite

- Aggiornamenti periodici per correggere vulnerabilità emergenti
- Log e auditing costante per rilevare anomalie e prevenire abusi
- Miglioramento continuo basato sui feedback e sulle analisi delle minacce reali

Le **contromisure** di sicurezza saranno parte **integrante** dell'**offerta delle soluzioni** AI/LLM di Fastweb ai clienti, garantendo così una protezione avanzata e affidabile.

AI Governance

Governance: abbiamo progettato la nostra AI Strategy attorno a 4 Pilastri

DATI

I dati sono **FONDAMENTALI** per un uso corretto dell' AI

Dobbiamo **GOVERNARE** i nostri dati, renderli **DISPONIBILI** e **AFFIDABILI**.

TECNOLOGIA

La nostra infrastruttura HPC **INSTALLATA E GESTITA IN ITALIA**

Le tecnologie AI devono diventare **PERVASIVE** all'interno dell'azienda

CASI D'USO

Ogni singolo contributo dei dipendenti alimenta l'**«AI AT SCALE»**

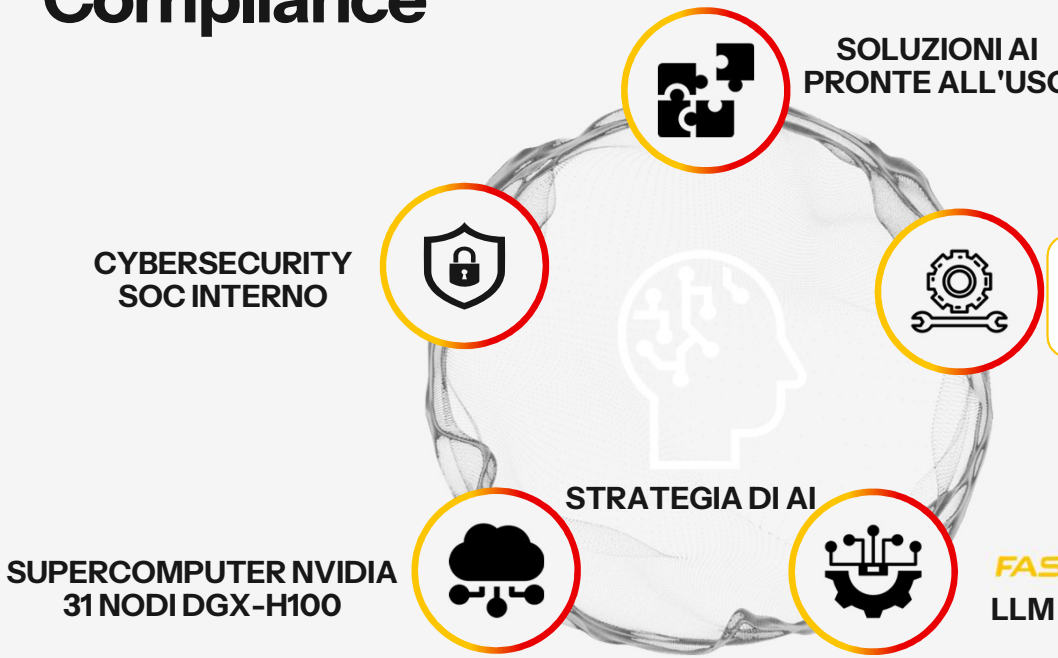
Le idee vengono valutate in base a **READINESS**, **ROI** e **CONTRIBUTO ALLA MATURITÀ DELL'AI**

COMPETENZE & ORGANIZZAZIONE

Modello **HUB AND SPOKE**: un COE AI centralizzato per sviluppare progetti di AI in forte collaborazione con i team aziendali.

I team possono diventare sempre più **INDIPENDENTI** man mano che la loro maturità nell'IA cresce

L'AI Governance supporta **la strategia** di adozione interna dell'Intelligenza Artificiale e garantisce gli **adempimenti di Compliance**



OBIETTIVO

- **Diffondere l'adoption dell'Intelligenza Artificiale**
- **In modo lecito, etico e human-centrico**
- **massimizzando il beneficio ed il ritorno sull'investimento**

Grazie all'adozione di questo modello di Governance AI

FASTWEB È MEMBRO DELL'AI PACT

(https://digital-strategy.ec.europa.eu/en/policies/ai-pact)

FASTWEB + vodafone

AI Compliance

AI Compliance – Come lo abbiamo implementato

MITIGAZIONE DEI RISCHI

- ✓ **Processo** e **strumenti** di analisi dei rischi
- ✓ **Valutazione dei rischi a 360°** (AI Act, GDPR, Copyright, ESG)
- ✓ **Misure** di mitigazione
- ✓ **Matrice** dei rischi

POLICY E CONTROLLI

- ✓ **Modello organizzativo** di AI - sfruttare le **sinergie di compliance**-
- ✓ **Codice di condotta** AI
- ✓ **Regole di sviluppo** dell'IA
- ✓ **Regole di trasparenza** dell'IA
- ✓ **Controlli** di I, II, III livello
- ✓ **Matrice** dei controlli

ORGANIZZAZIONE

- ✓ Assegnazione di **ruoli** e **responsabilità**
- ✓ Sviluppo dei **processi**
- ✓ **Flussi** di informazioni
- ✓ **Formazione** e **sensibilizzazione**

MONITORAGGIO CONTINUO

- ✓ **Monitoraggio** di normative, best practice e **provvedimenti delle autorità**
- ✓ **Aggiornamento** e/o **integrazione** degli strumenti aziendali

Conclusioni | Che cosa deve fare ogni realtà che voglia impiegare l'Intelligenza Artificiale?

Censimento dell'utilizzo di AI

Valutazione dei rischi legati all'AI

Formazione e Sensibilizzazione

Emissione di misure di mitigazione adeguate

Controllo della implementazione delle misure

Supervisione e monitoraggio del ciclo di vita dell'AI

Definizione dei ruoli, responsabilità e processi

Definizione delle regole e policy per l'adozione dell'AI

Grazie dell’attenzione!