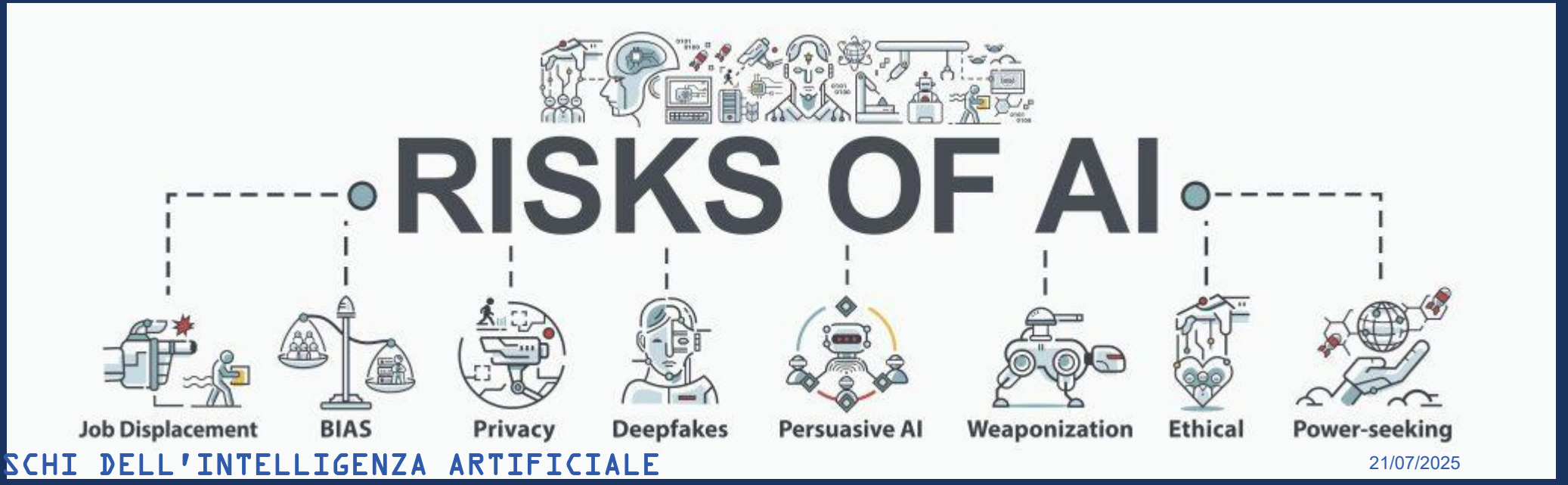
AI – INTRODUZIONE

L'INTELLIGENZA ARTIFICIALE AL SERVIZIO DELL'UOMO

I RISCHI DELL'INTELLIGENZA ARTIFICIALE

RISCHI DELL'USO INDISCRIMINATO DELL'INTELLIGENZA ARTIFICIALE

AMBIENTALI, SOCIALI, ETICI, TECNOLOGICI, COGNITIVI E CULTURALI



PERICOLI DELL'INTELLIGENZA ARTIFICIALE

- Perdita di Posti di Lavoro
- Impatto Ambientale
- Automazione della Guerra
- Superintelligenza
- Manipolazione dell'Opinione Pubblica
- Crimini Informatici

PERICOLI DELL'INTELLIGENZA ARTIFICIALE

Privacy e Sicurezza dei Dati

Raccolta ed elaborazione di grandi quantità di dati

Rischi significativi per la privacy e la sicurezza

Esempi: Furto di dati personali, uso improprio delle informazioni raccolte

Discriminazione e Bias

Algoritmi di IA possono perpetuare pregiudizi esistenti

RISCHI TECNOLOGICI E DI SICUREZZA (1)

🤖 Uso malevolo dell'IA:

- Creazione di deepfake, phishing o disinformazione automatizzata.

🔓 Vulnerabilità tecniche:

- Alcuni modelli IA possono essere ingannati con input manipolati.

- **Deepfake & Disinformazione**: contenuti audio/video manipolati possono distorcere informazioni, influenzare elezioni e diffondere truffe Wikipedia Reuters.
- **Cyber-attacchi avanzati**:
  - IA genera phishing, brute force, malware generativo, attacchi automatizzati; sviluppo di sistemi autonomi per la guerra cibernetica o militare.
  - Accesso non autorizzato a modelli AI ospitati su cloud o ambienti edge.
- **Perdita di controllo**: complessità e imprevedibilità dei modelli possono sfuggire al controllo umano AI Multiple WalkMe™ - Digital Adoption Platform.

RISCHI TECNOLOGICI E DI SICUREZZA (2)

🤖 Uso malevolo dell'IA:

- Creazione di deepfake, phishing o disinformazione automatizzata.

🔓 Vulnerabilità tecniche:

- Alcuni modelli IA possono essere ingannati con input manipolati.

- **Spiegabilità e trasparenza**: difficoltà nel comprendere le decisioni prese da modelli "black box" (es. deep learning).
- **Overfitting/Underfitting**: modelli troppo aderenti o troppo generici rispetto ai dati reali.
- **Allineamento dei valori**: l'IA può ottimizzare obiettivi in conflitto con valori etici umani Energy Central National Law Review.
- **Manipolazione dei dati / Data Poisoning** di input per ottenere risposte scorrette; corruzione intenzionale dei dati di training.

RISCHI PER LA PRIVACY E LA SORVEGLIANZA

- 👁️ Raccolta dati massiva:
  - Le IA apprendono da dati utente spesso senza consenso chiaro ➔ raccolta massiva di dati, tracciamento senza consenso .
- Vulnerabilità agli attacchi *(Adversarial AI)*:
  - Esfiltrazione di dati tramite AI inversa ➔ inferenza di dati sensibili dai modelli (model inversion attack).
- 🕵️ Sorveglianza automatizzata:
  - IA applicata a telecamere e social può monitorare persone e comportamenti.

RISCHI AMBIENTALI

🌡️ Consumo energetico e impatto climatico:

- L'addestramento di modelli IA (come GPT) richiede molta energia e produce CO_2.

🧯 Materiali critici:

- I datacenter richiedono hardware con terre rare, spesso estratte in modo non sostenibile.

- Consumo energetico massiccio: addestramento e utilizzo generano CO_2 comparabili a voli transcontinentali.
- Raffreddamento e acqua: i datacenter consumano enormi quantità d'acqua.
- Obsolescenza ed e-waste: aggiornamenti hardware più frequenti incrementano rifiuti elettronici .

RISCHI ETICI, LEGALI E SOCIALI (1)

⚠️ Discriminazione algoritmica:

- Se l'IA apprende da dati distorti, può penalizzare gruppi per etnia, genere o lingua *(es. esclusione di candidati da processi di selezione basati su attributi nascosti)*.
- **Bias algoritmico**: discriminazioni introdotte da dati di addestramento sbilanciati o da scelte progettuali errate.
- **Grading sociale automatizzato**: sistemi di reputazione e punteggio marginalizzano individui .
- **Pregiudizi incorporati dagli sviluppatori**: i programmatori trasmettono bias inconsci nel software .

DISCRIMINAZIONI ALGORITMICHE

Cosa sono: i bias sono pregiudizi insiti nei dati con cui vengono addestrati i modelli. Possono causare decisioni distorte su base etnica, di genere, socio-economica o altro.

📌 **Esempi concreti**:

- Un algoritmo sanitario che punta su spesa medica passata può discriminare le persone di colore Datatron.
- Software per l'assunzione penalizza donne o chi ha fatto maternità The Australian.
- Sistemi di riconoscimento facciale sbagliano fino al 35 % sui volti scuri .
- Il software COMPAS, usato nel sistema giudiziario USA, mostra bias razziali nei punteggi di recidività Wikipedia.

🔍 **Perché succede**:

- Dati di addestramento non rappresentativi (es. prevalenza di soggetti bianchi) .
- Assenza di trasparenza ("black box"), difficile controllare le decisioni .
- Preconcetti degli sviluppatori trasmessi involontariamente nei modelli .

RISCHI ETICI, LEGALI E SOCIALI (2)

- 💼 "Perdita" di lavoro:
  - L'automazione con IA può sostituire ruoli ripetitivi, aggravando le disuguaglianze ➔ chi controlla l'IA accumula vantaggio economico ➔ disoccupazione tecnologica

- 💼 Manipolazione sociale / opinione pubblica:
  - uso di AI per la generazione di contenuti (deepfake, bot) con fini propagandistici ➔ può influenzare comportamenti, opinioni, voti...

RISCHI ETICI, LEGALI E SOCIALI (3)

Rischi Normativi e Legali

- **Normativa incerta** ➔ regolamenti IA in evoluzione creano zone grigie giuridiche.
- **Assenza di accountability**: difficoltà nell'attribuire responsabilità in caso di danni o decisioni errate.
- **Violazione di normative**: es. GDPR (art. 22) in caso di decisioni automatizzate non spiegabili.
- **Conflitti giurisdizionali**: modelli AI distribuiti a livello globale che operano in contesti regolatori differenti.

RISCHI OPERATIVI E ORGANIZZATIVI

- Mancata governance del ciclo di vita dell'AI:
  - Modelli non monitorati dopo il deployment.
  - Assenza di policy per l'addestramento periodico.
- Scarsa qualità dei dati: dati errati o non aggiornati impattano negativamente sui risultati.
- Dipendenza tecnologica: lock-in da specifici vendor o framework AI.

RISCHI STRATEGICI E SISTEMICI

- AI generativa fuori controllo:
  - produzione automatica di codice, contenuti o decisioni con impatti imprevisti.
- Rischio esistenziale (Long-term AGI):
  - potenziale sviluppo di un'AI generale (AGI) non allineata agli obiettivi umani.
- Sbilanciamento geopolitico:
  - dominio di pochi attori globali sullo sviluppo dell'AI.

RISCHI EDUCATIVI E CULTURALI

📉 Appiattimento culturale:

- Le IA standardizzano le risposte, riducendo diversità di pensiero.
- Techno-solutionism: illusione che la tecnologia risolva ogni problema complesso
- Sostituzione di ruoli umani: l'IA può sostituire decisori come insegnanti, giudici, medici .

🛠️ Cultura della scorciatoia:

- Se lo studente si affida sempre all'IA, perde il gusto della ricerca autonoma.

RISCHI COGNITIVI E PSICOLOGICI

🧠 Dipendenza cognitiva:

- Affidarsi all'IA per tutto indebolisce la memoria e l'autonomia.

💣 Disinformazione:

- L'IA può generare risposte plausibili ma false, ingannando anche gli utenti più esperti.

DIPENDENZA COGNITIVA

Abitudine a chiedere tutto all'AI invece che riflettere da soli.

Cos'è: uso continuo e passivo dell'IA per decidere, generare idee, produrre testi → perdita dell'autonomia nel pensiero.

📌 **Esempio:** lo studente non si fida del proprio giudizio, chiede "se è giusto" a ChatGPT anche per cose che sa fare.

🔍 **Conseguenza:** difficoltà a prendere decisioni o a sviluppare un proprio stile espressivo.

SINTESI DEI RISCHI COGNITIVI

- Memoria → Non allenata, dipendenza dall'AI
- Immaginazione → Sostituita da output generati
- Pensiero critico → Ridotto per risposte preconfezionate
- Creatività → Omologazione dei contenuti
- Autonomia → Decisioni delegate all'AI
- Curiosità → Soddisfazione immediata senza ricerca

RIDUZIONE DELL'INTELLIGENZA OPERATIVA

Delegare all'AI riduce l'attività logica autonoma.

Cos'è: se deleghiamo all'IA operazioni come il calcolo, la scrittura, la pianificazione, rischiamo di usare meno le nostre capacità logico-razionali.

📌 **Esempio:** studenti che non esercitano più il problem solving perché usano l'IA per trovare subito la soluzione completa (non solo l'aiuto).

🔍 **Conseguenza:** minore esercizio del pensiero critico e della riflessione autonoma.

PERDITA DELLA FANTASIA

L'AI crea al posto nostro → rischiamo di non inventare più.
Cos'è: se lasciamo che sia l'IA a generare idee, storie, immagini, nomi, mappe concettuali, rischiamo di spegnere l'attività immaginativa umana.
📌 **Esempio:** "Fammi un racconto fantasy" → l'utente accetta passivamente la proposta dell'IA.
🔍 **Conseguenza:** indebolimento del pensiero divergente e della capacità di "inventare dal nulla".
🧩 **Suggerimento educativo:**
Usare l'IA come **stimolo iniziale**, ma poi **chiedere all'alunno di modificarla, migliorarla o sfidarla**.

ATROFIA DELLA MEMORIA

La facilità di accesso a risposte esterne indebolisce la memoria attiva.

Cos'è: la facilità con cui possiamo “riottenere tutto” ci disabitua a **ricordare concetti, testi, formule, passaggi logici**.

📌 **Esempio:** studenti che usano l’IA per trovare ogni volta definizioni, senza trattenerle a lungo termine.

🔍 **Conseguenza:** impoverimento della memoria semantica e perdita della struttura mentale della conoscenza.

🧩 **Suggerimento:**
Chiedere **di riprodurre una spiegazione** fatta da IA a parole proprie → aiuta a trasformare input in memoria attiva.

PERDITA DELLA CURIOSITÀ

Risposte rapide riducono la motivazione a cercare e approfondire.

Cos'è: se ogni risposta è a portata di prompt, viene meno il piacere della ricerca, dell'approfondimento, della scoperta graduale.

📌 **Esempio:** lo studente fa copia-incolla di una sintesi invece di leggere e farsi domande personali.

🔍 **Conseguenza:** impoverimento della motivazione intrinseca e della capacità di porsi domande.

ECCESSIVA FIDUCIA IN RISPOSTE SBAGLIATE

L'AI può informazioni false ➔ «allucinazioni».

Cos'è: i modelli AI possono generare risposte **coerenti ma false**. Chi li usa passivamente può **credere in contenuti errati**.

📌 **Esempio:** ChatGPT inventa una fonte o confonde concetti storici → lo studente li assume come corretti.

🔍 **Conseguenza:** difficoltà nel discernere tra informazione e disinformazione.

SVALUTAZIONE DELLO SFORZO

Se tutto è perfetto e immediato, viene meno il valore dell'errore.

Cos'è: l'IA offre risposte perfette e senza errori apparenti → lo studente si abitua a **non fallire** e perde il valore dell'apprendimento per tentativi.

📌 **Esempio:** evitare di riscrivere un testo perché "tanto me lo fa l'AI".
🔍 **Conseguenza:** perdita della resilienza cognitiva, minore tolleranza alla fatica.

COSA RISCHIAMO SE DELEGHIAMO TUTTO ALL'IA?

Facoltà umana	Rischio con uso passivo dell'IA
Memoria	Non allenata, si appoggia sempre all'esterno
Immaginazione	Sostituita da output IA
Pensiero critico	Atrofizzato da risposte pronte
Creatività	Omologata ai pattern appresi dall'IA
Autonomia decisionale	Affidata a una "macchina"
Curiosità	Sostituita da soddisfazione immediata

COME AFFRONTARLI?

- ✅ Educazione al pensiero critico:
  - Chiedere agli studenti di verificare le fonti e riflettere sui contenuti.

- 🔍 Etica e moderazione:
  - Limitare l'uso automatico, usare l'IA come **supporto**, non come **sostituto**.

SOLUZIONI PER MITIGARE I RISCHI DELL'IA

- Gestione del Rischio
  - Implementare un framework di gestione del rischio
  - Utilizzo di strumenti di monitoraggio continuo
- Regolamentazione e Normative
- Bias e Equità
- Sicurezza dei Dati
- Trasparenza e Interpretabilità

SOLUZIONI PER MITIGARE I RISCHI DELL'IA

- Supervisione Umana
- Formazione e Consapevolezza
- Manutenzione e Aggiornamenti
- Valutazione dell'Impatto
- Collaborazione e Condivisione delle Conoscenze

GESTIONE DEL RISCHIO NELL'IA

Identificazione dei Rischi

Riconoscere e documentare tutte le potenziali fonti di rischio

Esempi: rischi legati ai dati, rischi operativi, rischi etici

Analisi dei Rischi

Valutare la probabilità e l'impatto di ciascun rischio

Esempi: utilizzo di matrici di rischio

GESTIONE DEL RISCHIO NELL'IA

Valutazione dei Rischi

Prioritizzare i rischi in base alla loro analisi

Esempi: focalizzarsi sui rischi ad alta probabilità e alto impatto

Trattamento dei Rischi

Monitoraggio e Revisione

Comunicazione e Consultazione

STRUMENTI E FRAMEWORK PER LA GESTIONE DEL RISCHIO NELL'IA

Framework di Gestione del Rischio

Strutture formali che guidano l'intero processo di gestione del rischio

ISO/IEC 23894:2023 e ISO 31000 come esempi di standard internazionali

Valutazione del Rischio (Risk Assessment)

Processo di identificazione, analisi e valutazione dei rischi specifici associati a un sistema di IA

Identificazione delle vulnerabilità nei dati di addestramento

Analisi delle potenziali minacce alla sicurezza

Valutazione dell'Impatto (Impact Assessment)

Analisi degli effetti specifici che i sistemi di IA possono avere su individui, gruppi sociali o la società

Valutazione dell'impatto sulla privacy

Analisi delle conseguenze etiche

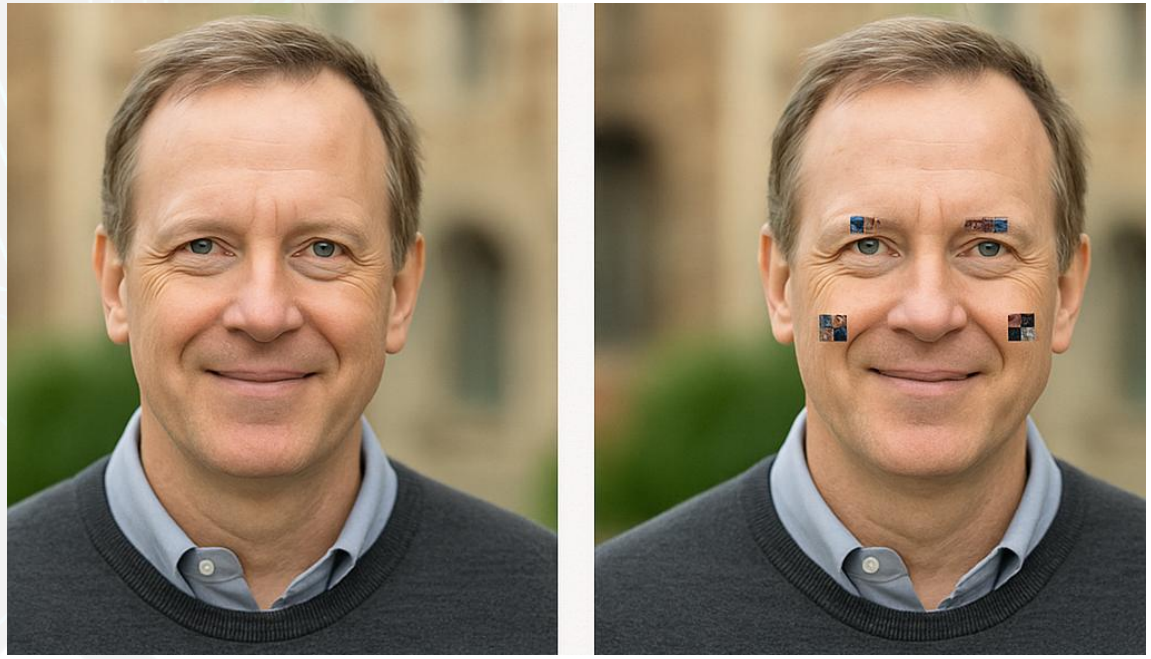
Strumenti di Monitoraggio e Reporting

ESEMPI DI RISCHI E MITIGAZIONI

Rischi Relativi ai Dati	Mitigazione: Implementare misure di sicurezza avanzate come la crittografia e l'anonimizzazione dei dati Esempi: Protezione dei dati sensibili durante l'addestramento e l'implementazione dei modelli di IA
Rischi Relativi al Modello	Mitigazione: Utilizzare tecniche di interpretabilità e trasparenza per comprendere meglio il funzionamento dei modelli di IA Esempi: Implementazione di tecniche di spiegazione come LIME (Local Interpretable Model-agnostic Explanations)
Rischi Operativi	Mitigazione: Effettuare manutenzione regolare e aggiornamenti dei sistemi di IA per garantire prestazioni ottimali Esempi: Monitoraggio continuo delle prestazioni dei modelli e aggiornamenti basati sui nuovi dati
Rischi Etici e Legali	Cercare una Normativa comune

1. RISCHI TECNOLOGICI – ESEMPI CONCRETI

- Amazon (2018): algoritmo HR penalizzava automaticamente i CV femminili.
- Adversarial AI: piccoli disturbi trasformano un cartello STOP in 'limite 45 km/h'.
- Data poisoning: 50 immagini modificate hanno alterato un sistema di riconoscimento facciale.



2. RISCHI ETICI E SOCIALI – ESEMPI CONCRETI

- COMPAS (USA): bias razziale nel rischio di recidiva.
- Clearview AI: scraping di milioni di immagini senza consenso dai social.
- Deepfake politici: video falsi di Zelensky, Biden e altri usati per propaganda.

3. RISCHI DI SICUREZZA INFORMATICA – ESEMPI CONCRETI

- Phishing generato da AI: email personalizzate scritte da LLM.
- Model inversion: recupero di dati sensibili da un modello AI.
- Attacchi AI-to-AI: bot maligni che causano danni interagendo con chatbot aziendali.

4. RISCHI NORMATIVI E LEGALI – ESEMPI CONCRETI

- Violazione GDPR art. 22: rifiuto automatico di un prestito senza spiegazione.
- App medica USA in UE: priva di marcatura CE, in violazione della MDR.
- Tesla Autopilot: responsabilità non chiara in caso di incidente.

5. RISCHI OPERATIVI E ORGANIZZATIVI – ESEMPI CONCRETI

- AI logistica post-COVID: previsioni sbagliate per dati obsoleti.
- Sentiment analysis obsoleta: non interpreta slang ed emoticon moderne.
- Vendor lock-in: azienda bloccata su Azure OpenAI, senza possibilità di migrazione.

Problema dopo il COVID-19:

- Cambiamenti radicali e improvvisi nel comportamento dei consumatori:
  - picchi imprevisti su prodotti sanitari (mascherine, igienizzanti),
  - crolli nella domanda di abbigliamento da ufficio o viaggi,
  - boom dell'e-commerce alimentare.
- Modelli AI addestrati su dati pre-pandemia (2017–2019) hanno continuato a:
  - sottostimare nuove tendenze,
  - sovrastimare prodotti obsoleti,
  - fallire nel ricalibrare tempi di consegna durante crisi logistiche (es. blocco del Canale di Suez).
- Assenza di retraining adattivo o real-time learning ha aggravato la situazione.

6. RISCHI STRATEGICI E SISTEMICI – ESEMPI CONCRETI

- Copilot: suggerisce codice vulnerabile, es. SQL injection.
- Sovranità UE: dipendenza da AI USA/Cina, a scapito della regolazione europea.
- AGI mal allineata: scenario teorico di superintelligenza pericolosa.

GRAZIE