





# Modulo 6: Etica e Responsabilità nell'NLP

Un viaggio tra opportunità e  
responsabilità

## Cosa tratteremo oggi

- Introduzione all'etica nell'NLP
- Bias e fairness nei sistemi NLP
- Privacy e sicurezza dei dati
- Trasparenza e spiegabilità
- Impatto sociale e responsabilità
- Casi di studio etici nell'NLP
- Framework etici e linee guida
- Sviluppo responsabile di sistemi NLP

## Introduzione all'etica nell'NLP

- L'NLP media sempre più le nostre interazioni quotidiane 
- I sistemi NLP **non sono strumenti neutri** 
- Incorporano valori, priorità e visioni del mondo 
- Prendono decisioni che un tempo erano esclusivamente umane 

"Con grande potere derivano grandi responsabilità"

# Definizione di Algoretica

L'algoretica è un campo emergente dell'etica che si occupa di studiare e valutare i problemi morali legati allo sviluppo, all'uso e all'impatto degli algoritmi, dei dati e delle pratiche correlate, specialmente nell'ambito dell'intelligenza artificiale (IA) e del Natural Language Processing (NLP).

Questa disciplina si basa su valori condivisi e principi morali che guidano il comportamento nella società, andando oltre le leggi per concentrarsi su ciò che è considerato "giusto" o "sbagliato" in contesti tecnologici

L'algoretica applicata al NLP richiede particolare attenzione, poiché le tecnologie linguistiche influenzano molti aspetti della vita quotidiana e possono perpetuare bias o causare danni se non gestite correttamente.

The urgency of an algorethics

# Definizione di Algoretica

Di seguito alcuni principi chiave e domande etiche da considerare:

- **Contesto della raccolta dati:** Il contesto in cui i dati sono stati raccolti corrisponde al contesto del loro utilizzo? Ad esempio, utilizzare dati raccolti per scopi accademici in un'applicazione commerciale potrebbe violare aspettative etiche
- **Incentivi e bias nella raccolta dati:** I dati sono stati raccolti da persone o sistemi con quote o strutture di incentivazione che potrebbero introdurre distorsioni?
- **Rappresentatività:** Chi è sottorappresentato o assente nei dati utilizzati per addestrare modelli di NLP? È possibile trovare dati aggiuntivi o utilizzare metodi statistici per rendere i dataset più inclusivi?
- **Consenso e scelte significative:** I dati sono stati raccolti in un ambiente in cui i soggetti avevano scelte significative e consapevoli riguardo alla loro partecipazione?

# Perché l'etica nell'NLP è importante? 🔍

- Adozione in contesti **critici e sensibili** 🏥👤
  - Sanità, giustizia, istruzione, finanza
- Impatto su **decisioni significative** che influenzano vite reali 🧬
- Potenziale di **amplificare disuguaglianze** esistenti 📈
- Necessità di bilanciare **innovazione e protezione** 🛡️

## Domanda 🤔

**Quali sistemi NLP utilizzate quotidianamente?**

**Avete mai notato comportamenti problematici?**




# Bias e fairness nei sistemi NLP ⚖️

## Origini dei bias:

- **Dati di addestramento** 📊
  - Sottorappresentazione di gruppi
  - Rappresentazioni stereotipate
- **Scelte di progettazione** 🛠️
  - Feature, definizione del problema, metriche
- **Contesto di utilizzo** 🌍
  - Disparità di accesso e performance



# Manifestazioni di bias nei sistemi NLP

- **Word embeddings biased** 
  - "uomo : programmatore :: donna : casalinga"
- **Generazione di testo discriminatoria** 
  - Descrizioni stereotipate di certi gruppi
- **Classificazione iniqua** 
  - Performance degradata per lingue minoritarie



# Misurazione e mitigazione dei bias 📏

## Metriche di fairness:

- Demographic parity, Equal opportunity, Equal accuracy




## Tecniche di mitigazione:

- Interventi sui dati 🖌️
  - Bilanciamento, data augmentation
- Interventi algoritmici 🎲
  - Debiasing di embeddings, adversarial learning
- Interventi post-processing 🔄
  - Calibrazione, re-ranking, filtering






# Privacy e sicurezza dei dati

## Sfide uniche nell'NLP:

- **Informazioni personali nel testo** 
  - PII, informazioni sensibili, comportamentali
- **Memorizzazione nei modelli** 
  - Riproduzione verbatim, inferenza da memorizzazione
- **Inferenze non autorizzate** 
  - Profilazione demografica e psicografica

# Tecniche per la privacy-preserving NLP

- **Anonimizzazione e de-identificazione** 
  - NER per PII, redaction, pseudonimizzazione
- **Privacy differenziale** 
  - Aggiunta di rumore controllato
- **Federated learning** 
  - Addestramento locale sui dispositivi




## Quiz per la platea!

**Quale di queste NON è una tecnica di privacy-preserving NLP?**

- A) Federated learning
- B) Differential privacy
- C) Gradient boosting
- D) Pseudonimizzazione

# Sicurezza e attacchi ai sistemi NLP

## Vulnerabilità specifiche:

- **Adversarial attacks** 
  - Text perturbations, prompt injection, jailbreaking
- **Data poisoning** 
  - Inserimento di esempi malevoli nei dati
  - Backdoor attacks
- **Model stealing** 
  - Estrazione di modelli proprietari



# Framework normativi e compliance 📄

- **Regolamentazioni rilevanti** ⚖️
  - GDPR (EU), CCPA/CPRA (California), HIPAA (US)
- **Principi di privacy by design** 🏗️
  - Minimizzazione dei dati
  - Limitazione dello scopo
  - Privacy come impostazione predefinita
  - Trasparenza

# Trasparenza e spiegabilità 🔍

## L'importanza della trasparenza:

- **Trasparenza sui dati** 📊
  - Provenienza, caratteristiche, annotazione
- **Trasparenza algoritmica** 📈
  - Architettura, iperparametri, metriche
- **Trasparenza operativa** ⚙️
  - Scopo, processo decisionale, supervisione

# Sfide alla spiegabilità nell'NLP 🤖

- **Complessità e opacità** 🧩
  - Miliardi di parametri
  - Rappresentazioni distribuite
  - Comportamenti emergenti
- **Trade-off tra performance e spiegabilità** ⚖️
  - Modelli più potenti tendono ad essere più opachi

# Tecniche per l'interpretabilità nell'NLP

## Approcci principali:

- **Interpretabilità intrinseca** 
  - Attention visualization
  - Sparse models
- **Interpretabilità post-hoc** 
  - LIME, SHAP, Feature attribution
- **Spiegazioni in linguaggio naturale** 
  - Rationale generation
  - Counterfactual explanations

## **Momento di riflessione**

**Se un sistema NLP prende una decisione importante che vi riguarda...**


**Quali informazioni vorreste avere sul suo funzionamento?**

# Impatto sociale e responsabilità 🌍

## Impatto multidimensionale:

- **Accesso all'informazione e filter bubbles** 🔍
  - Algoritmi che mediano l'accesso all'informazione
- **Disinformazione e manipolazione** 📰
  - Generazione di fake news, deepfake testuali
- **Impatto sul lavoro e sull'economia** 💼
  - Automazione di compiti cognitivi
- **Impatto su lingue e culture** 🗣️
  - Disparità linguistiche, omogeneizzazione culturale

## **Impatto sul lavoro**

- Automazione di mansioni ripetitive
- Nuove opportunità (traduzione, assistenza)
- Rischio sostituzione lavoro umano 

### **Domanda:**

Quali lavori potrebbero sparire? Quali nuovi lavori nasceranno?

# Responsabilità degli sviluppatori e delle organizzazioni

## Approcci proattivi:

- **Valutazione dell'impatto** 
  - Impact assessment, stakeholder engagement
- **Design responsabile** 
  - Value-sensitive design, inclusive design
- **Governance e accountability** 
  - Chiara attribuzione di responsabilità
  - Meccanismi di oversight



# Una domanda provocatoria 🤔

**Chi è responsabile quando un sistema NLP causa un danno?**

- Lo sviluppatore del modello?
- L'organizzazione che lo implementa?
- L'utente che lo utilizza?
- Il regolatore che non ha imposto limiti adeguati?





# Casi di studio etici nell'NLP 📖

## Esempi emblematici:



- **Bias di genere nei sistemi di traduzione** 🌐
  - Traduzioni stereotipate da lingue gender-neutral
- **Moderazione dei contenuti e libertà di espressione** 🗣️
  - Impatto sproporzionato su comunità marginalizzate
- **Privacy nei modelli linguistici di grandi dimensioni** 🔒
  - Memorizzazione e potenziale rivelazione di dati personali

# Framework etici e linee guida

## Principi etici fondamentali:

- **Beneficenza e non maleficenza** 
  - Massimizzare benefici, minimizzare danni
- **Autonomia e consenso informato** 
  - Rispetto delle scelte individuali
- **Giustizia ed equità** 
  - Distribuzione equa di benefici e rischi
- **Trasparenza e accountability** 
  - Apertura su funzionamento e responsabilità

# Framework etici esistenti

- Framework istituzionali 
  - [Principi AI dell'OECD](#)
  - [Ethics Guidelines for Trustworthy AI \(EU\)](#)
- Framework industriali 
  - [Microsoft Responsible AI Principles](#)
  - [Google AI Principles](#)

# Implementazione pratica dei framework etici

## Strumenti e processi:




- **Strumenti di valutazione etica** 
  - Ethical impact assessment, Ethics checklists
- **Processi di governance** 
  - Ethics review boards, Ethics by design
- **Cultura organizzativa** 
  - Leadership commitment, Incentivi allineati
  - Diversità e inclusione

# Sviluppo responsabile di sistemi NLP 🌱

## Ethics by design:

- **Fase di concezione e pianificazione** 🧩
  - Valutazione della necessità, definizione di scopo e limiti
- **Fase di raccolta e preparazione dei dati** 📊
  - Sourcing etico, valutazione di rappresentatività
- **Fase di sviluppo del modello** 🛠️
  - Scelte di architetture, monitoraggio durante l'addestramento
- **Fase di testing e valutazione** 🔍
  - Test multidimensionali, adversarial testing
- **Fase di deployment e monitoraggio** 🚀
  - Deployment graduale, monitoraggio continuo

# Strumenti pratici per lo sviluppo responsabile

- **Documentazione standardizzata** 
  - Datasheets for Datasets
  - Model Cards
- **Toolkit e librerie** 
  - Fairness Indicators
  - What-If Tool
  - AI Fairness 360
- **Processi e framework** 
  - Responsible AI Maturity Model
  - Ethics Canvas

# Collaborazione multidisciplinare







- **Team multidisciplinari** 
  - Esperti di etica, scienze sociali, legge, domain experts
- **Coinvolgimento degli stakeholder** 
  - Participatory design
  - Community engagement
- **Formazione e sensibilizzazione** 
  - Curriculum integration
  - Leadership awareness




## Riflessione finale

**L'etica nell'NLP non è un ostacolo all'innovazione, ma una componente essenziale di un'innovazione veramente benefica e sostenibile.**

## Conclusione

- I sistemi NLP incorporano valori e priorità con profonde conseguenze sociali 
- I bias possono perpetuare o amplificare disuguaglianze esistenti 
- La privacy è particolarmente critica per dati linguistici personali 
- La trasparenza e la spiegabilità sono essenziali per costruire fiducia 
- Lo sviluppo responsabile richiede un approccio multidisciplinare 
- L'etica nell'NLP è un processo continuo, non una destinazione finale 

## Sfide future

- Maggiore explainability
- Modelli più inclusivi
- Regolamentazione internazionale 

### Domanda:

Chi dovrebbe decidere le regole? Governi, aziende, utenti?

## ✨ Come ci ricorda Spider-Man

“Da grandi poteri derivano grandi responsabilità” 🕷️

💬 **Domanda finale:**

Come usereste il vostro ‘potere NLP’ per fare del bene?

## Riferimenti e Approfondimenti

- Bender, E. M., et al. (2021). **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?**
- Blodgett, S. L., et al. (2020). **Language (Technology) is Power: A Critical Survey of "Bias" in NLP**
- Floridi, L., & Cowls, J. (2019). **A Unified Framework of Five Principles for AI in Society**
- Weidinger, L., et al. (2021). **Ethical and social risks of harm from Language Models**

 **Grazie per l'attenzione!**

 Domande? Commenti?

Pronti a cambiare il mondo... in modo etico!