UNIVERSITÀ POLITECNICA DELLE MARCHE

DII
Dipartimento di Ingegneria dell'Informazione

unimc

# Intelligenza Artificiale e Attacchi Cognitivi.


Enrico Mercogliano.

Dipartimento: Ingegneria dell'Informatica.

Università Politecnica delle Marche.

Contatti: enricone84@gmail.com


Giovedì 2 - Venerdì 3 Ottobre 2025

Corso di Perfezionamento in Cybersecurity, Cyber Risk and Data Protection

# Introduzione.

In questa presentazione si esamina come modelli di **Intelligenza Artificiale (IA)** possono essere utilizzati per condurre **attacchi cognitivi**, ovvero influenzare subdolamente i processi mentali e le decisioni di individui e gruppi attraverso contenuti persuasivi o manipolativi. Adottando una prospettiva **interdisciplinare** (informatica, scienze cognitive, comunicazione, psicologia, diritto, geopolitica, economia), verranno analizzati sia gli strumenti tecnologici (modelli IA per generare e diffondere contenuti) sia le **tecniche di manipolazione** basate su tali strumenti, e infine gli **impatti socio-cognitivi** che ne derivano.

# Obiettivi.

- Comprendere i principali modelli di IA utilizzati per creare testi, immagini, video e altri media **persuasivi** o **ingannevoli**, e il loro funzionamento di base.
- Riconoscere diverse **tecniche di attacco cognitivo** (es. disinformazione automatizzata, deepfake, manipolazione del comportamento via microtargeting psicografico) e valutare casi di studio reali su scala internazionale.
- Analizzare gli **effetti** di questi attacchi sulla società e sulla mente umana: manipolazione dell'opinione pubblica, distorsione dei processi democratici, erosione della fiducia nelle istituzioni, polarizzazione sociale, impatti sulla salute mentale e **distorsioni nell'economia dell'informazione**.
- Individuare collegamenti tra prospettive diverse (es. implicazioni legali dei deepfake, basi cognitive della persuasione, dimensioni geopolitiche della disinformazione, incentivi economici alla diffusione di fake news) e sviluppare un approccio critico e multidisciplinare al problema.

# Parte 1: Modelli AI per contenuti persuasivi/manipolativi

# Modelli di linguaggio.

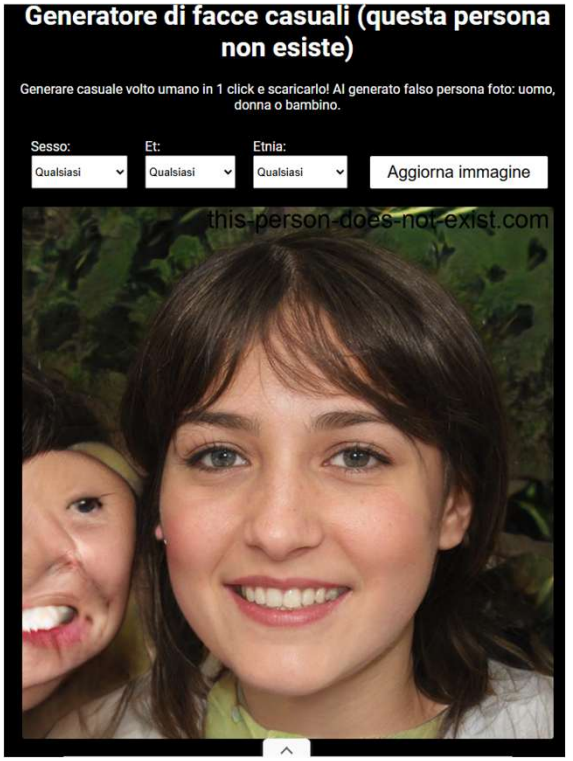
NLP (*Natural Language Processing*): reti neurali di grandi dimensioni (es. **Transformer** e modelli autoregressivi tipo *GPT-3/4*) in grado di produrre testi coerenti e di alta qualità. Questi sistemi possono generare articoli, post sui social o messaggi personalizzati su larga scala, **imitando stili comunicativi umani**.

*Esempi:* OpenAI GPT-3 è stato addestrato su enormi quantità di dati testuali ed è capace di scrivere articoli convincenti; ricercatori hanno dimostrato che GPT-3 può produrre **propaganda politica efficace**: in un esperimento, i lettori esposti a testi generati dall'IA hanno concordato con tesi propagandistiche nel 43% dei casi, quasi quanto con propaganda umana (47%), e superando quest'ultima (53%) se l'IA veniva guidata e curata da esseri umani.

Questo dimostra il **potenziale persuasivo dei modelli linguistici**, già sfruttato da alcuni governi (es. segnalazioni di **post social generati da IA legati al governo cinese** per influenzare elettori a Taiwan e negli USA.

# Modelli generativi di immagini e video.

Algoritmi di **visione artificiale** come le **reti antagoniste generative (GAN)** o modelli di diffusione, capaci di creare volti e scene fittizie altamente realistici. Questi sono alla base dei **deepfake** video (volti sovrapposti per far apparire qualcuno mentre dice o fa cose mai accadute) e di immagini false utilizzate in contesti manipolativi.



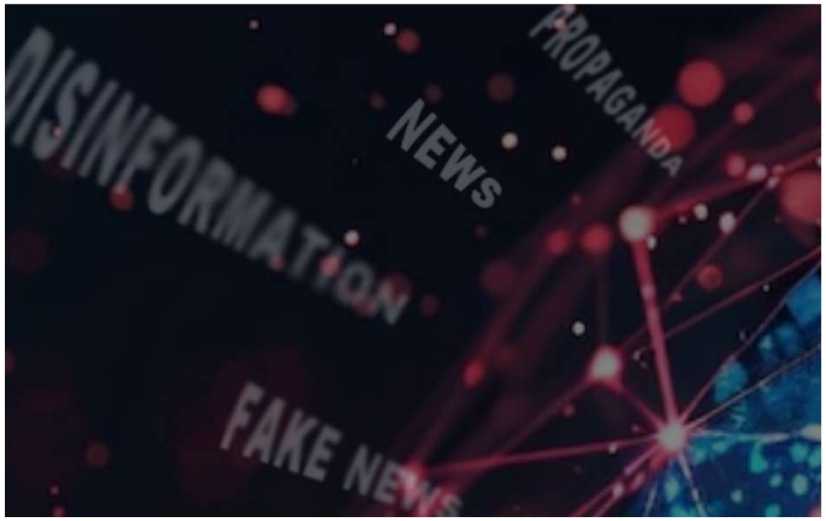
# Sistemi di raccomandazione e algoritmi di diffusione.

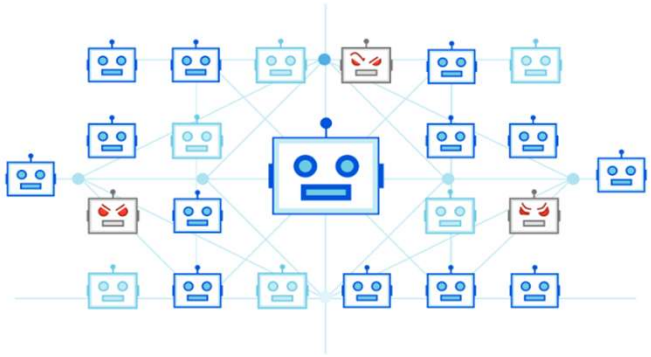
Sebbene non generino contenuti, gli algoritmi di piattaforme come Facebook, YouTube, TikTok **influenzano fortemente la diffusione** e la visibilità dei messaggi. Basati su **machine learning** (profilazione degli interessi, ottimizzazione dell'engagement), essi possono creare **"bolle informative"** e amplificare contenuti emozionali o controversi che massimizzano i clic, spesso a scapito dell'accuratezza.


# Parte 2: Tecniche di Attacco Cognitivo

# Disinformazione automatizzata e bot sociali.

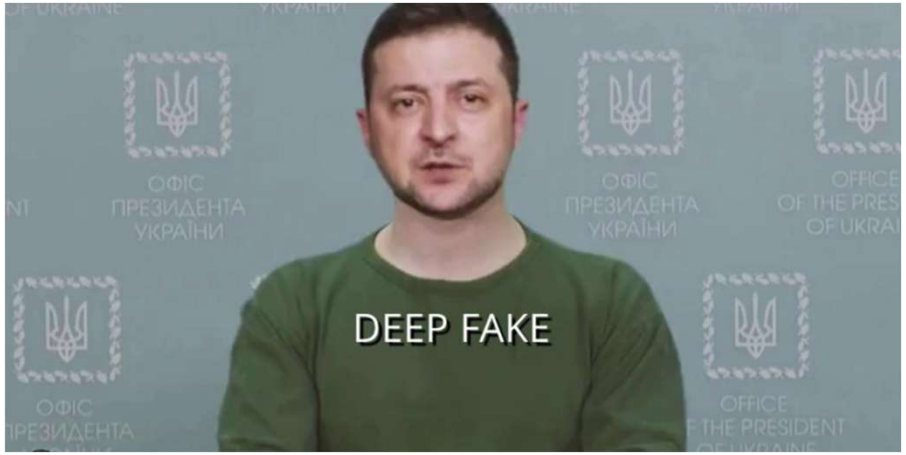
**Disinformazione automatizzata e coordinata,** si tratta della diffusione di grandi volumi di notizie false o fuorvianti in modo **automatizzato**, spesso tramite reti di bot o account falsi, con l'obiettivo di *plagiare l'opinione pubblica* o confondere i factchecker.

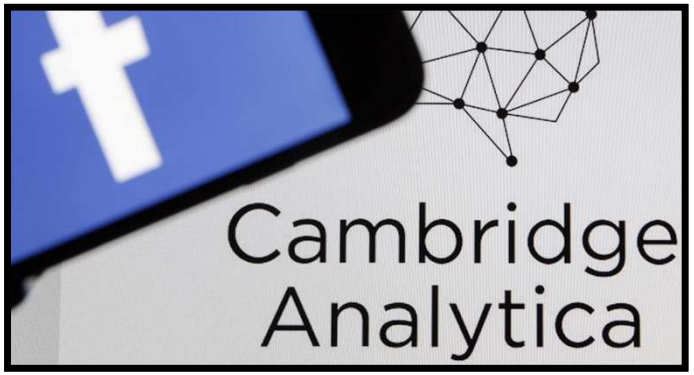
# Deepfake e media sintetici.

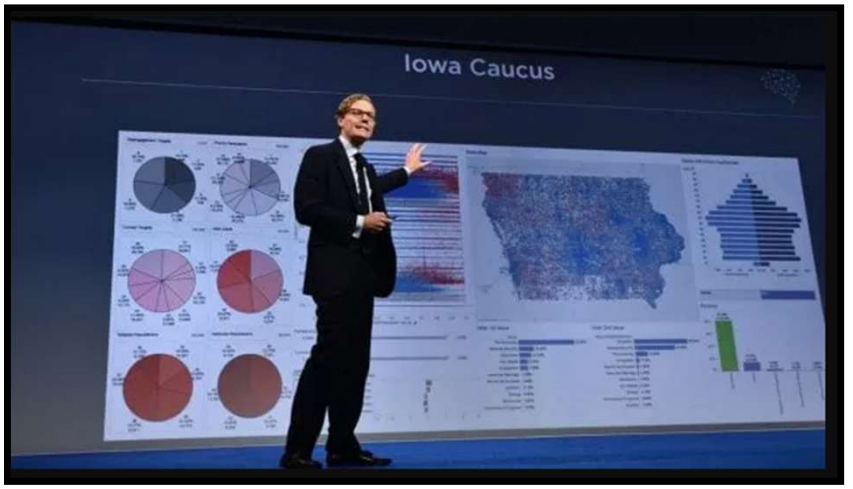
**Deepfake e media sintetici** sono video/audio falsificati per ingannare il pubblico, screditare persone o diffondere false dichiarazioni.


# Microtargeting psicografico e manipolazione comportamentale via IA.

È una tecnica che unisce big data, psicologia e IA per *profilare* individui e gruppi in base a tratti psicologici (es. personalità, valori, paure) e inviare messaggi su misura per manipolarne percezioni e comportamenti. Caso emblematico: **Cambridge Analytica** (USA 2016 & Brexit 2016) società anglo-americana che ha raccolto senza autorizzazione i dati personali di **decine di milioni di utenti Facebook** (fino a 87 milioni) tramite un quiz online, costruendo modelli di personalità dettagliati.



# Altre tecniche e scenari.

*Operazioni psyop in contesti geopolitici* sono attacchi cognitivi supportati da IA che rientrino in strategie più ampie di **“guerra cognitiva”**. Un esempio la propaganda online dell’ISIS per il reclutamento; sebbene non basata su IA avanzata, percorre l’uso strumentale dei social per manipolare convinzioni e attirare foreign fighters; il caso **Myanmar 2021**, dove un video apparso (forse manipolato) mostrava un ex ministro detenuto “confessare” corruzione dell’ex leader Aung San Suu Kyi, usato dalla giunta militare per giustificare il golpe, molti osservatori sospettarono un deepfake, segno che queste tattiche stanno entrando anche in regimi autoritari di altre regioni. Si menzioneranno anche le **“troll farms”** cinesi impegnate a influenzare la percezione internazionale, es. campagne pro-Pechino su X (ex-Twitter) con account automatizzati, utilizzando volti generati per profili fake analoghe a quelle russe già viste.

# Parte 3: Impatti Sociali e Cognitivi, Risposte e Dibattito

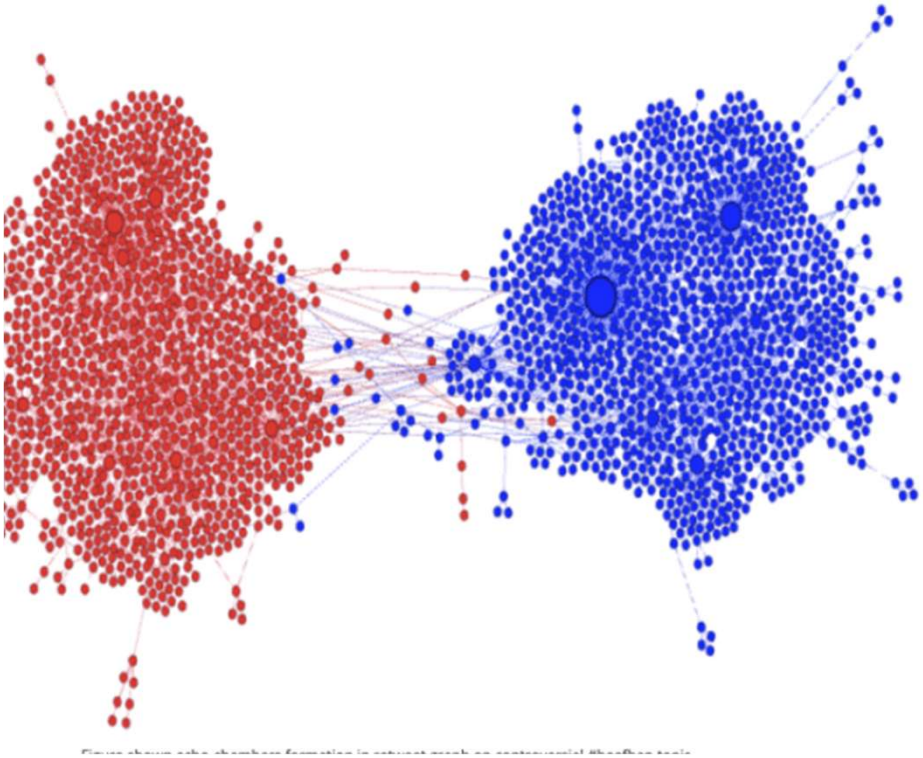
# Erosione dell’opinione pubblica informata e distorsione dei processi democratici

Le campagne di disinformazione e manipolazione massiva possono alterare la formazione dell’opinione pubblica, creando *false percezioni di realtà*. Un esempio lampante è la diffusione della **“Big Lie”** (la falsa teoria di brogli nelle elezioni USA 2020)

# Crisi di fiducia nelle istituzioni e nei media tradizionali

L’effetto cumulativo di campagne di disinformazione e attacchi cognitivi è spesso la **delegittimazione dell’autorità**. Quando si diffondono costantemente teorie del complotto o deepfake che coinvolgono leader, l’audience può sviluppare cinismo e **dietrologia generalizzata** “non credo più a nulla”.

# Polarizzazione sociale ed “echo chambers"



Come evidenziato in precedenza, i meccanismi algoritmici e le campagne manipolative contribuiscono a dividere l’opinione pubblica in fazioni chiuse. Questo porta a fenomeni di *polarizzazione estrema*: non solo disaccordo politico, ma vera e propria **settarizzazione** (tribalismo “noi vs loro”). Uno studio pubblicato su *Science* (2020) ha definito i social media un fattore che intensifica il “settarianismo politico” negli USA.

# Effetti sulla salute mentale e sul comportamento individuale

L’esposizione prolungata a disinformazione e contenuti manipolativi può avere impatti psicologici significativi. Ad esempio, “**infodemia**” (overflow di notizie, spesso false) ha generato in molti **ansia, stress e confusione**. A livello micro, subire continuamente messaggi distorti può portare a **cognitive overload** (sovraccarico informativo) e riduzione della capacità di concentrazione e pensiero critico. Sul piano emotivo, teorie cospirazioniste martellanti possono indurre sentimenti di paranoia o sfiducia generalizzata verso il prossimo, incidendo sul benessere mentale. Si para anche del fenomeno delle **molestie amplificate dall’IA**: deepfake *pornografici* o campagne diffamatorie automatizzate possono colpire individui, causando traumi psicologici, paura e ritiro dalla vita pubblica (soprattutto per donne e minoranze bersaglio di tali attacchi).

# Economia dell’informazione e costi socio-economici

Gli attacchi cognitivi basati su IA incidono anche sull’economia dei media e sulla società in termini di risorse. Il **modello pubblicitario online** premia l’attenzione: “**click = denaro**”. Questo crea un ecosistema dove spesso la *disinformazione sensazionalistica è economicamente più redditizia* dell’informazione di qualità. Ciò mina l’**industria dell’informazione** tradizionale (che fatica a competere con il clickbait virale) e contribuisce alla chiusura di testate locali e al taglio di giornalisti, indebolendo l’ecosistema informativo verificato. Inoltre, la **lotta alla disinformazione** ha costi elevati: le piattaforme investono miliardi in moderazione di contenuti e algoritmi di detection, i governi devono spendere in alfabetizzazione mediatica e unità anti-fake news, i partiti politici devono contrastare campagne tossiche, tutte risorse sottratte ad altro. In termini macro, la **sfiducia generata** ha impatto economico. Un altro aspetto è la possibilità di **manipolazioni di mercato**: sono già avvenuti episodi in cui false notizie (talvolta propagate da bot) hanno mosso i mercati finanziari o crypto, creando volatilità e perdite, scenario destinato a peggiorare se un domani un deepfake convincente di un CEO che annuncia il falso fallimento di una compagnia circolasse prima della smentita.