UNIVERSITÀ
POLITECNICA
DELLE MARCHE

DII
Dipartimento di Ingegneria
dell'Informazione

unIMC

Disinformazione ed IA: un approccio multidisciplinare

Emanuele Paltrinieri

Giovedì 2 Ottobre 2025

Corso di Perfezionamento in Cybersecurity, Cyber Risk and Data Protection

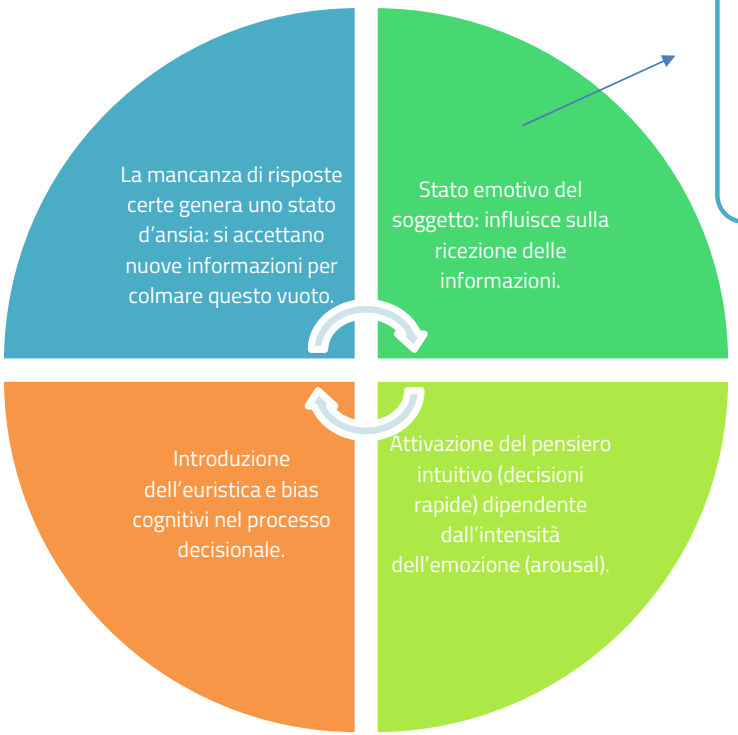
Distorsione delle informazioni

Distorsione delle informazioni

Mal-information, mis-information e dis-information: definizioni

Mal-information	Mis-information	Dis-information
▪ Contenuto vero ▪ Intenzione generalmente (ma non sempre) lesiva/ ingannevole	▪ Contenuto falso/fuorviante ▪ Priva di intenzione lesiva/ingannevole	▪ Contenuto falso/fuorviante ▪ Intenzione volutamente lesiva/ingannevole

Disinformazione: aspetti neurologici e psicologici



- La rabbia spinge il soggetto a rafforzare le proprie convinzioni e a rifiutare informazioni ritenute incongruenti.

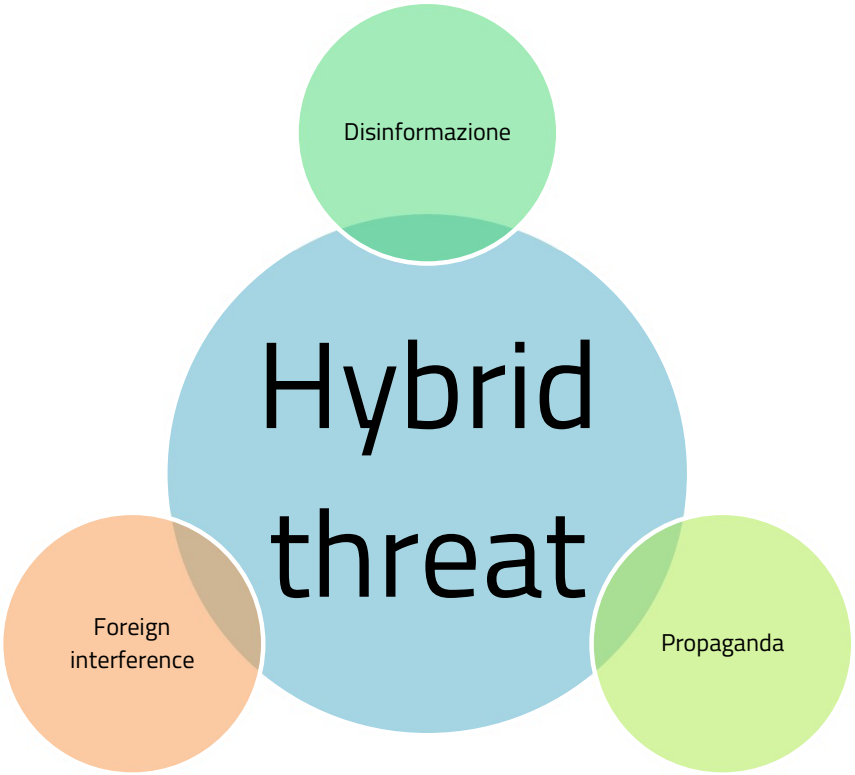
Disinformazione, propaganda e Foreign interference

Foreign interference (FI)

Foreign interference

- Viene definita come «qualsiasi interferenza illegittima da parte di potenze straniere nei processi politici e democratici dell'UE e dei suoi stati membri». (Mildebrath, 2024)
- Include: «pratiche di manipolazione online, finanziamenti o campagne di finanziamenti illeciti a partiti, operazioni di influenza, attacchi cyber a infrastrutture elettorali e azioni dirette contro individui». (Mildebrath, 2024)

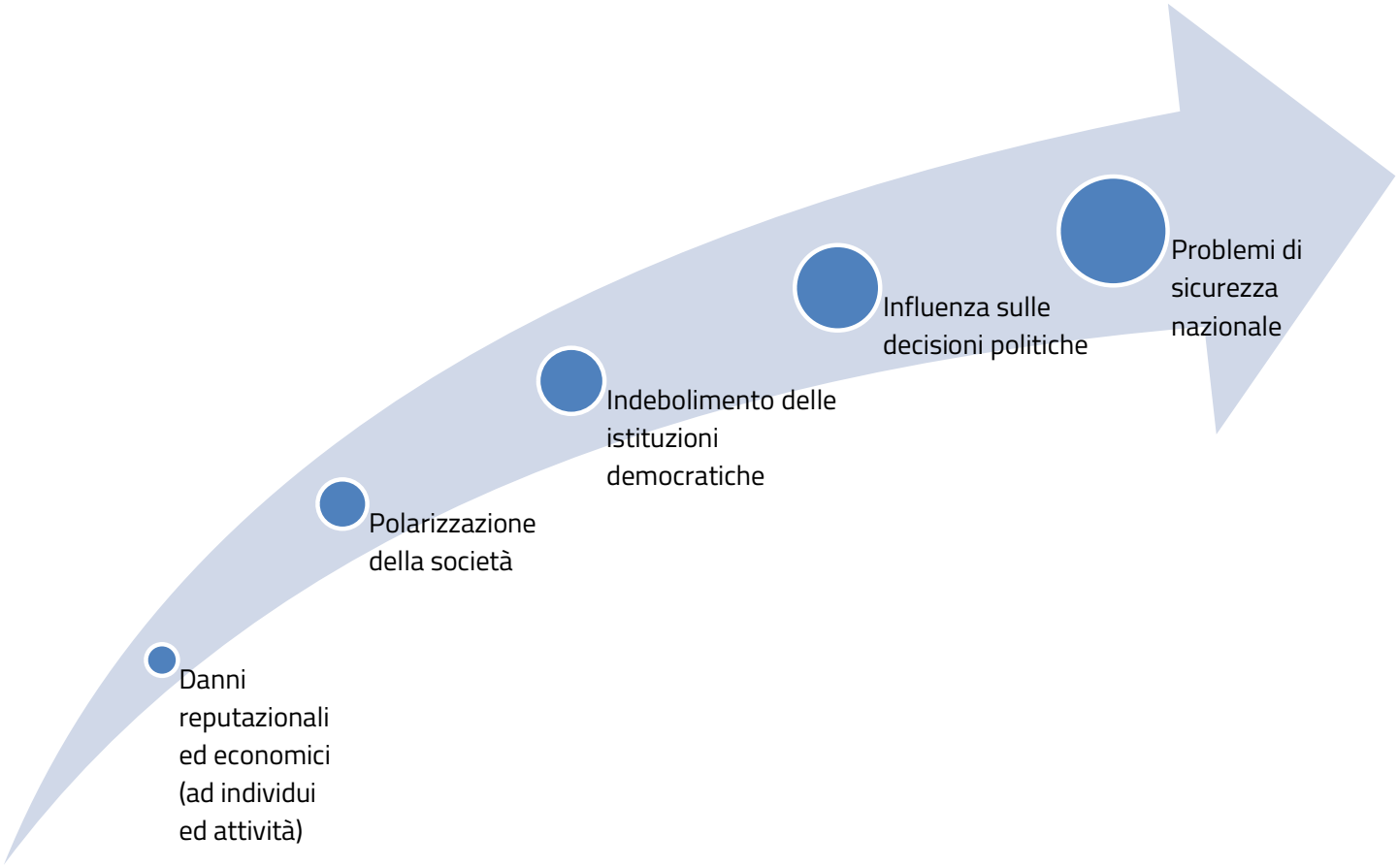
Hybrid threat



Hybrid threat

- Conventional and unconventional methods (diplomatic, military, economic, technological), which can be used in a coordinated manner by state or non-state actors to achieve specific objectives while remaining below the threshold of formally declared warfare. (Tsaruk, Knorniiets, 2020, p. 70)

Rischi


Danni reputazionali ed economici (ad individui ed attività)
Polarizzazione della società
Indebolimento delle istituzioni democratiche
Influenza sulle decisioni politiche
Problemi di sicurezza nazionale

Vantaggi della disinformazione per l'attaccante

Economico

- Costi di attuazioni contenuti rispetto alle strategie convenzionali.

Competenze

- Discreta conoscenza della tecnologia, ma non necessariamente tecnico-specialistica.
  Ampliamento della platea di attori con intenti malevoli.

Proiezione

- Ogni individuo/istituzione può diventare un target, indipendentemente dall'area geografica.

Efficacia

- I social ne potenziano la propagazione e gli effetti; favoriscono l'anonimato.

Alcuni dati: attori, incidenti e piattaforme

Repubblica popolare cinese

Federazione russa

Corea del Nord

Fonte: 3rd EEAS Report on Foreign Information Manipulation and Interference Threats, EEAS, 2025, p. 6. https://www.eeas.europa.eu.

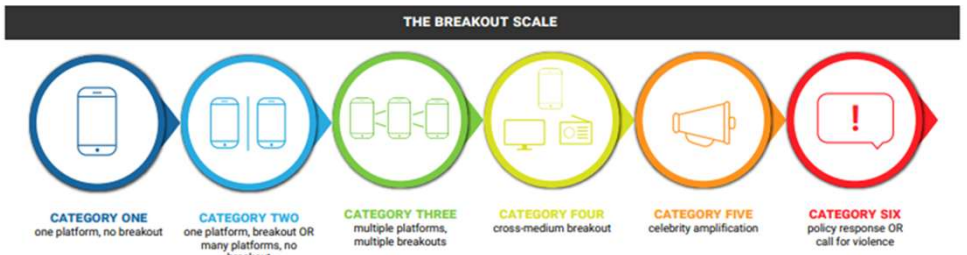
Strumenti: individuazione e misurazione dell'impatto

1. Framework ABCDE (Actors, Behaviour, Content, Degree, Effect) (Pamment, 2020, p. 6);
2. Break-out scale: misura il livello di propagazione di un' operazione di influenza sulle piattaforme online, attribuendo una scala di valori da 1 a 6. (Nimmo, 2020, p. 6)

The ABCDE Framework

Actor	*What kinds of actors are involved?* This question can help establish, for example, whether the case involves a foreign state actor.
Behavior	*What activities are exhibited?* This inquiry can help establish, for instance, evidence of coordination and inauthenticity.
Content	*What kinds of content are being created and distributed?* This line of questioning can help establish, for example, whether the information being deployed is deceptive.
Degree	*What is the overall impact of the case and whom does it affect?* This question can help establish the actual harms and severity of the case.
Effect	*What is the overall impact of the case and whom does it affect?* This question can help establish the actual harms and severity of the case.

Fonte: Nimmo, 2020, p. 6



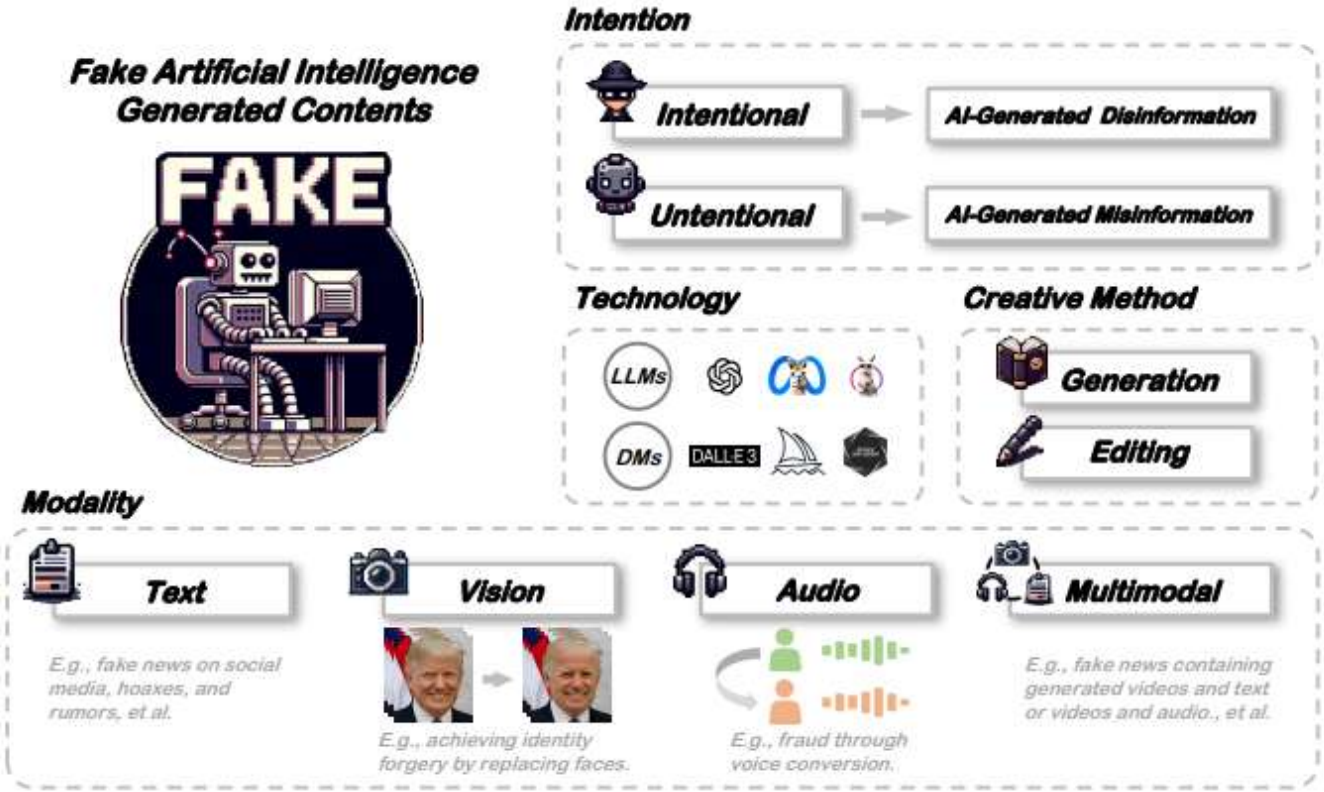
Fonte: Pamment, 2020, p. 6

Intelligenza artificiale (IA) e disinformazione

Impieghi dell'IA

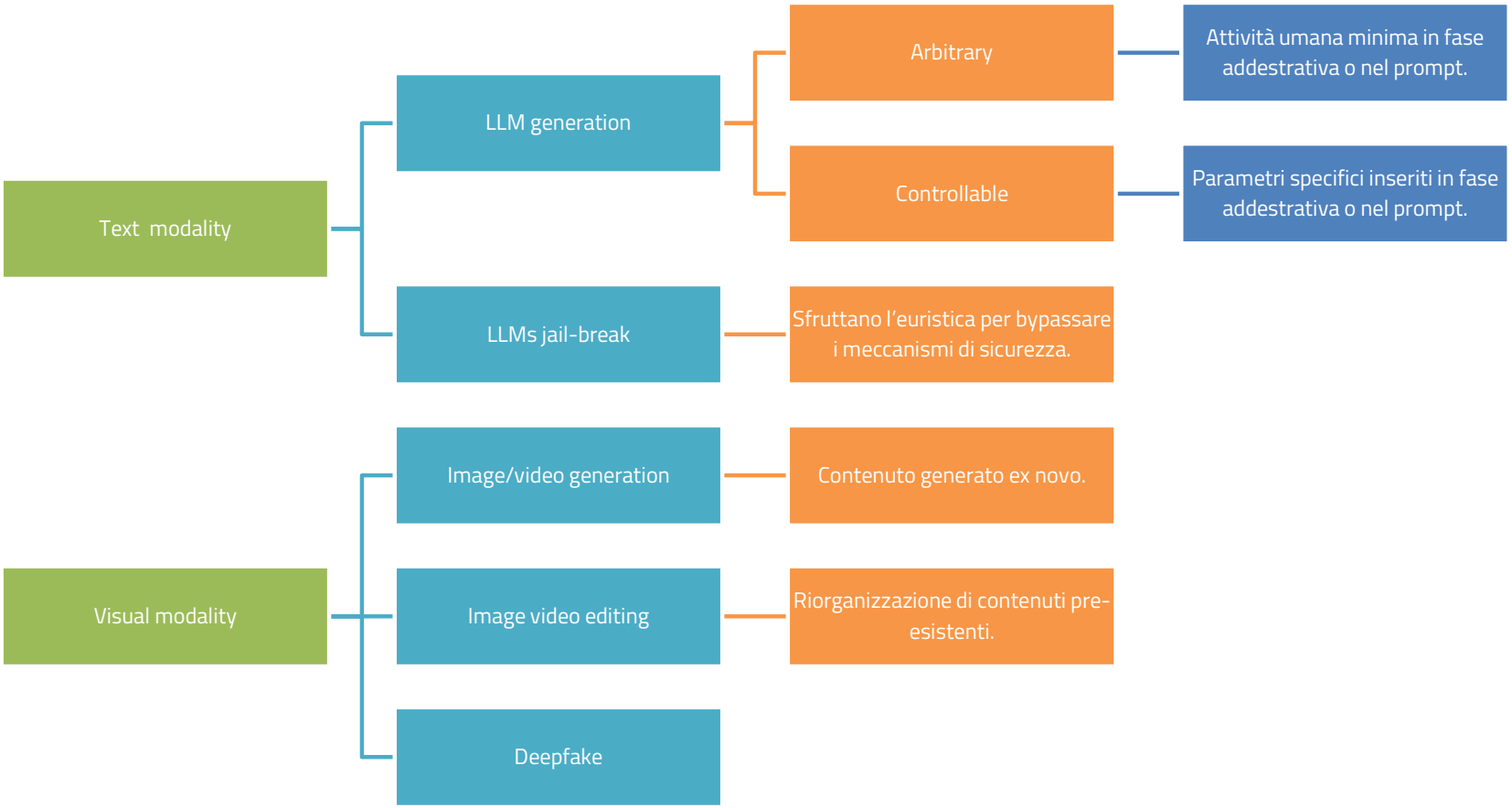
- l'IA trova impiego in diversi ambiti (commerciale, marketing, dubbing, assistenti virtuali), tra i quali la creazione di contenuti informativi.
- Agisce come un «φάρμακον» (nel mondo antico, il termine indicava sia un veleno sia il suo rimedio), ovvero può avere un effetto nocivo o benefico in base all'utilizzo.
- Nello specifico, può essere impiegata:
  - Nella creazione/editing di contenuti falsi sfruttando l'IA generativa;
  - Per avvelenare i dati (data poisoning) in fase di addestramento (introduzione dati errati o obsoleti, introduzione di bias).

Fake Artificial Intelligence Generated Contents (FAIGC): una panoramica



Fonte: Xiaomin Yua, Yezhaohui Wanga, Yanfang Chen, Zhen Tao, Dinghao Xi, Shichao Song, Simin Niu, Zhiyu Lia, p. 4.

Focus: Text modality e Visual modality



Deep-fake: metodi

Metodi

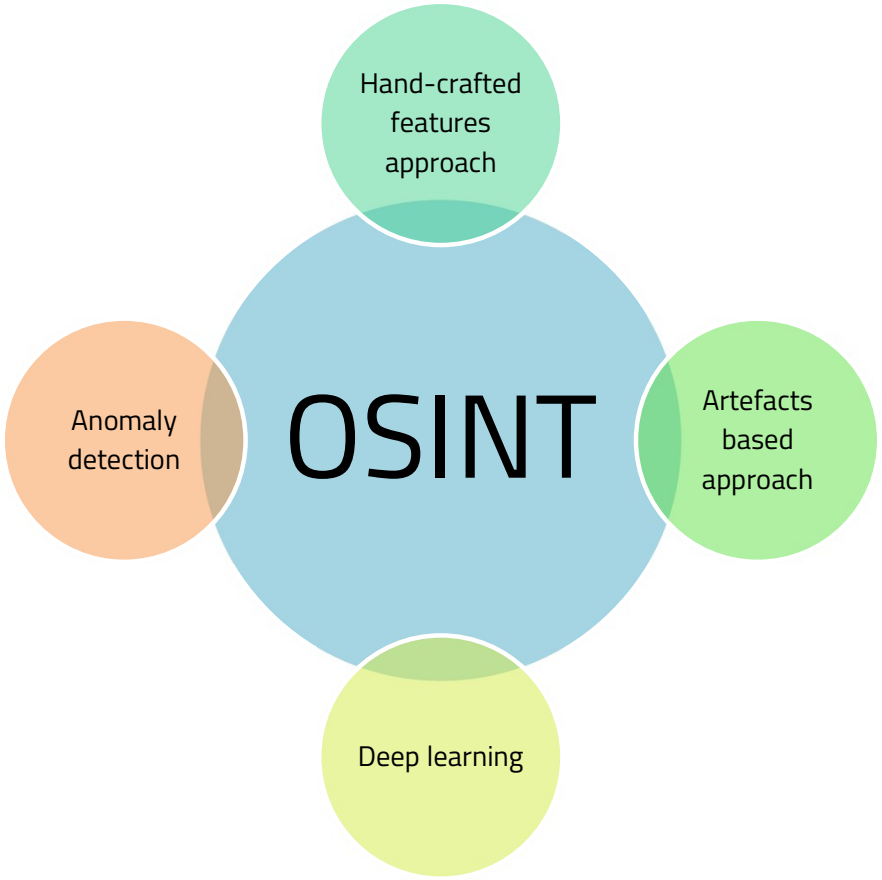
- Creazione di un volto inesistente;
- Identity swap: un volto viene sostituito con un altro;
- Attribute manipulation: vengono modificati i tratti di un volto;
- Expression swap: viene sostituita l'espressione di un volto con un'altra.
- Face reenactement (più innovativa): trasferimento di un'espressione da un video ad un'immagine target. Nel caso si tratti di persone diverse, si parla di cross-reenactement.

Deep fake: tecniche

Tecniche

- Auto-encoders: l'encoder prende un input e lo trasforma in una rappresentazione numerica (spazio latente), operando una sintesi e compressione dei dati, mentre il decoder svolge l'operazione inversa;
- GANs: il principio è quello degli auto-encoders, ma si addestra il modello creando una competizione tra generatore e discriminatore;
- Latent space decomposition: rappresentazione dei dati che cattura solo le componenti essenziali;
- Diffusion models: sfrutta la distribuzione della probabilità dei dati in una dimensione inferiore, cattura gli elementi dello spazio latente inserendoli in una gerarchia.

Deep-fake: metodi di contrasto



Hand-crafted features

- Analisi statistica dei pixel per studiare le modifiche.

Artefacts based approach

- Analisi dei parametri umani modificati e confronto con quelli reali (movimento delle labbre, espressioni…).

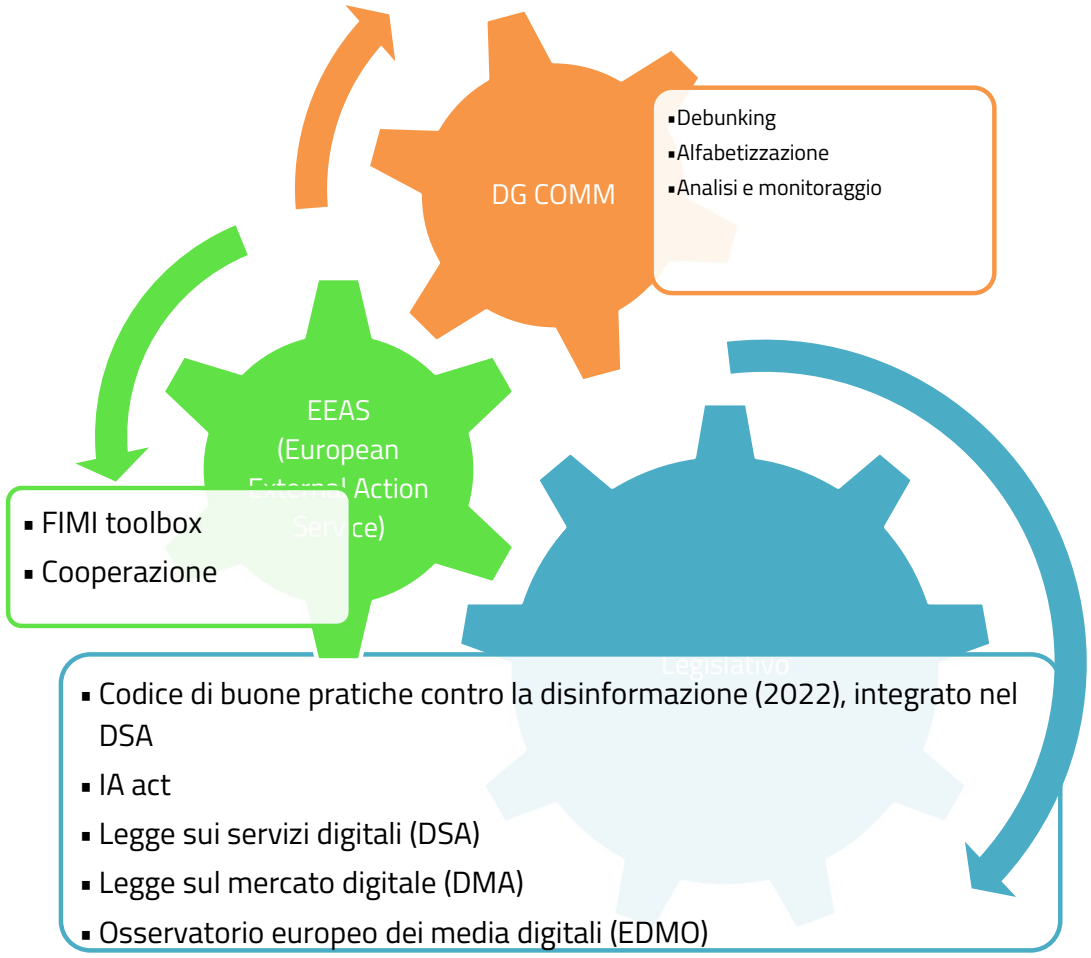
Deep-learning

- Si analizza il fake sfruttando la tecnica della classificazione.

Anomaly detection

- Si analizza il fake cercando anomalie rispetto ad una media ed operando una classificazione tra input reali e falsi.

Unione Europea: strumenti per il contrasto



Obiettivi in sintesi:

I. Rimozione tempestiva dei contenuti illeciti, inclusi disinformazione e *hate speech* (DSA, DMA);
II. Riduzione operazioni di influenza (DSA);
III. Maggiore tutela dei minori (DSA, DMA);
IV. API potenziata per l'analisi e la prevenzione (DSA);
V. Sistema di sanzioni (comune);
VI. Obblighi di *due diligence* (comune);
VII. Maggiore trasparenza (comune);
VIII. Maggiore coordinazione e collaborazione tra istituzioni (comune);
IX. Maggiore alfabetizzazione digitale (comune);

Bibliografia

- Bodenhausen, G. V., Sheppard, L. A., & Kramer, G. P. (1994). Negative affect and social judgment: The differential impact of anger and sadness. European Journal of Social Psychology, 24(1), 45–62. https://doi.org/10.1002/ejsp. 2420240104
- Boyer, M. M. (2021). Aroused argumentation: How the news exacerbates motivated reasoning. The International Journal of Press/politics. https://doi. Org/10.1177/1940161221101057
- Cherry, S. (2024). Modern armed conflicts: Disinformation campaigns shaping the digital information landscape. Serials Librarian, 85(1–4), 19–31).
- DeSteno, D., Petty, R. E., Rucker, D. D., Wegener, D. T., & Braverman, J. (2004). Discrete emotions and persuasion: The role of emotion-induced expectancies. Journal of Personality and Social Psychology, 86(1), 43–56. https://doi. org/10.1037/0022-3514.86.1.43
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology, 1(1), 13–29. https://doi.org/10.1038/ s44159-021-00006-y
- Freiling, I., Krause, N. M., Scheufele, D. A., & Brossard, D. (2021). Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. New Media & Society. https:// doi.org/10.1177/14614448211011451
- Greenstein, M., & Franklin, N. (2020). Anger increases susceptibility to misinformation. Experimental Psychology, 67(3), 202–209. https://doi.org/10.1027/ 1618-3169/a000489
- Hasell, A., & Weeks, B. E. (2016). Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media: partisan news, emotions, and information sharing. Human Communication Research, 42(4), 641–661. https://doi.org/10.1111/hcre.12092
- Hendrick Mildebrath, Member's Research Service PE 760.355, March 2024.

Bibliografia

- Kim, H., Park, K., & Schwarz, N. (2010). Will this trip really be exciting? The role of incidental emotions in product evaluation. Journal of Consumer Research, 36(6), 983–991. https://doi.org/10.1086/644763
- Kreps, S., McCain, R. M., & Brundage, M. (2020). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. Journal of Experimental Political Science, 9(1), 104–117.
- Lühring, J., Shetty, A., Koschmieder, C. et al. Emotions in misinformation studies: distinguishing affective state from emotional response and misinformation recognition from acceptance. Cogn. Research 9, 82 (2024). https://doi.org/10.1186/s41235-024-00607-0
- MacKuen, M., Marcus, G., Neuman, W. R., & Miller, P. R. (2010). Affective Intelligence or Personality? State vs. Trait Infuences on Citizens' Use of Political Information (SSRN Scholarly Paper 1643468). https://papers.ssrn.com/ abstract=1643468
- Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. Cognitive Research: Principles and Implications, 5, 1–20. https://doi.org/10.1186/s41235-020-00252-3
- Hendrick Mildebrath, Member's Research Service PE 760.355, March 2024.
- Moreno Espinosa, P., Abdulsalam Alsarayreh, R. A., & Figuereo-Benítez, J. C. (2024). Big data and artificial intelligence as solutions to disinformation. Doxa Comunicacion, 38.
- Nimmo B., The Breakout Scale: measuring the impact of influence operations, Brookings, 2020.
- Oleksandr TSARUK & Maria KORNIIETS, 2020. "**Hybrid nature of modern threats for cybersecurity and information security**," Smart Cities and Regional Development (SCRD) Journal, Smart-EDU Hub, Faculty of Public Administration, National University of Political Studies & Public Administration, vol. 4(1), pages 57-78, March.
- Pamment, J. (2020). The Organization of the EU's Disinformation Policy, in The EU's Role in Fighting Disinformation: Crafting A Disinformation Framework (pp. 10–17). Carnegie Endowment for International Peace. http://www.jstor.org/stable/resrep26180.7
- Salaverría, R., Bachmann, I., & Magallón-Rosa, R. (2024). Desinformación y confianza en los medios: Propuestas de actuación. Comunicación, 14(2), 13–32.

Bibliografia

- 3rd EEAS Report on Foreign Information Manipulation and Interference Threats, EEAS, 2025, https://www.eeas.europa.eu. [Visitato il 14 agosto 2025]
- Tharindu F. , Darshana Priyasad, Sridha Sridharan, Arun Ross, Fookes C. Face Deepfakes - A Comprehensive Review Cornell University, Thu, 13 Feb 2025 23:08:05 UTC, https://doi.org/10.48550/arXiv.2502.09812 [Visitato il 14 agosto 2025]
- Xiaomin Yua, Yezhaohui Wanga, Yanfang Chen, Zhen Tao, Dinghao Xi, Shichao Song, Simin Niu, Zhiyu Lia, Fake Artificial Intelligence Generated Contents (FAIGC): A Survey of Theories, Detection Methods, and Opportunities, Cornell University, Fri, 3 May 2024 04:47:01 UTC, https://doi.org/10.48550/arXiv.2405.00711. [Visitato il 14 agosto 2025]