ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Laurea Magistrale in Governance e Politiche dell'Innovazione Digitale

SICUREZZA COGNITIVA: DISINFORMAZIONE, IA GENERATIVA E RESILIENZA DEMOCRATICA NELL'ERA DELLE MINACCE IBRIDE

GOVERNANCE DELLA CYBERSECURITY (B5812)
A.A. 2025/26
Susanna Cioni - mat. 1164027
Laura Tonsi - mat. 1164043

ABSTRACT

La diffusione di tecniche di manipolazione informativa abilitate dalla *Generative AI* rende la dimensione cognitiva un obiettivo strategico delle operazioni ibride contemporanee. L'automazione della persuasione, la scalabilità della disinformazione e la capacità di incidere sui processi decisionali sotto-soglia introducono una vulnerabilità sistemica non affrontabile con strumenti tecnici o regolatori isolati.

Il presente lavoro propone un modello di governance volto a rafforzare la sicurezza cognitiva attraverso l'integrazione di tre componenti: orientamento normativo, metodologie di valutazione del rischio algoritmico e strumenti di difesa attiva. L'analisi è stata sviluppata combinando fonti esistenti con una procedura esplorativa basata su simulazione dialogica tramite modelli generativi (LLM), utilizzati per testare le soluzioni disponibili, evidenziarne i limiti e valutare la resilienza di configurazioni alternative.

Il risultato è un framework che, pur non offrendo una soluzione definitiva, fornisce una struttura operativa utile a coordinare interventi oggi frammentati e a orientare la governance verso la protezione dell'integrità epistemica. Il lavoro evidenzia inoltre aree ancora aperte — dalla misurazione dell'impatto cognitivo alle implicazioni giuridiche della tutela epistemica — indicando la necessità di sviluppi interdisciplinari futuri.

CAPITOLO 1 - INTRODUZIONE: L'ECOSISTEMA DEL DISORDINE INFORMATIVO E LA NUOVA FRONTIERA DELLA SICUREZZA

1.1 Il cambio di paradigma: Dalla Cyber Security alla Cognitive Security

La trasformazione del *cyberspazio* nell'ultimo decennio ha imposto una ridefinizione radicale del perimetro di sicurezza per le organizzazioni pubbliche e private. Tradizionalmente, la *cybersecurity* si è fondata sulla protezione dell'infrastruttura logica e fisica (server, reti, *endpoint*), garantendo la confidenzialità, l'integrità e la disponibilità dei dati (la classica triade CIA - *Confidentiality, Integrity and Availability).* In questo contesto, emerge il concetto di *Cognitive Security* (Sicurezza Cognitiva), definita come la protezione dei processi mentali e decisionali da manipolazioni esterne intenzionali. Come evidenziato dai rapporti del *NATO Strategic Communications Centre of Excellence*, le moderne operazioni ibride mirano a sfruttare le vulnerabilità della psicologia umana piuttosto che i *bug* del *software*. L'attaccante non mira più necessariamente a disabilitare un sistema informatico (*Denial of Service*), ma ad alterare la percezione della realtà di chi lo governa (*Denial of Reality*), inducendo errori strategici, panico finanziario o polarizzazione sociale.[1]

Per gli apparati di sicurezza nazionale e le istituzioni democratiche, ignorare questa dimensione significa esporsi a un rischio sistemico. Una campagna di disinformazione ben orchestrata può causare l'erosione della fiducia nelle istituzioni, la polarizzazione sociale estrema e la paralisi dei processi decisionali critici, con impatti che spesso superano quelli di un attacco cinetico tradizionale. Di conseguenza, la *governance* della sicurezza deve passare da un'ottica esclusivamente tecnico-difensiva a un approccio globale che congiunga la protezione delle infrastrutture di rete alla salvaguardia dell'integrità informativa.[2]

1.2 Oltre le "Fake News": Tassonomia del Disordine Informativo

Per affrontare il problema con rigore metodologico, è imperativo superare l'uso del termine "*fake news*". Tale espressione, divenuta politicamente carica e semanticamente vaga, è stata giudicata inadeguata dalla comunità scientifica internazionale per descrivere la complessità delle minacce attuali.[3]

La tassonomia proposta da Wardle e Derakhshan (2017) nel rapporto fondamentale *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, commissionato dal Consiglio d'Europa, è universalmente riconosciuta come il riferimento principale per i legislatori e i *risk manager*. Gli autori argomentano che l'informazione dannosa deve essere classificata incrociando due dimensioni: la falsità del contenuto e l'intenzione di nuocere.

---

[1] **NATO StratCom COE.** (2021). *Social media manipulation 2021: State of the art*. NATO Strategic Communications Centre of Excellence.
[2] **Transparency International.** (2024). *Fake news, corruption and compliance in the private sector*. Transparency International Helpdesk, p. 2.
[3] **Wardle, C., & Derakhshan, H.** (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe Report, p. 5.

Ne derivano tre categorie distinte, ognuna delle quali richiede specifiche strategie di mitigazione:

1. ***Misinformation*** **(Disinformazione involontaria):** si verifica quando un'informazione falsa viene condivisa senza l'intenzione di causare danno. Un esempio tipico si ha quando, durante una nuova crisi (come un disastro naturale o un conflitto in corso), vengono involontariamente condivise sui social media vecchie foto, video di repertorio (*old footage*) o filmati tratti da crisi precedenti o da altri contesti, spacciandoli per contenuti in tempo reale. Altri esempi includono la semplice negligenza nel riportare statistiche o dati inaccurati.
2. ***Disinformation*** **(Disinformazione intenzionale):** è la creazione e condivisione consapevole di informazioni false con l'intento specifico di causare danno a una persona, un gruppo sociale, un'organizzazione o uno Stato. Rientrano in questa categoria le campagne di propaganda statale, le operazioni di *influence elettorale* e le frodi finanziarie basate su notizie fabbricate.
3. ***Malinformation*** **(Informazione malevola):** costituisce forse la sfida più insidiosa. Si tratta di informazioni autentiche (o basate sulla realtà) usate intenzionalmente per infliggere danno, spesso spostate da un contesto privato o sensibile a quello pubblico. Esempi tipici includono il *doxxing* (cioè, la divulgazione di dati personali senza consenso), il *revenge porn* o le *fughe di notizie* strategiche relative a comunicazioni interne di organizzazioni sensibili, come si è verificato nei casi Sony o durante le elezioni USA del 2016.

Questa distinzione è cruciale per la Governance. Mentre la *Misinformation* può essere contrastata con l'educazione e la correzione (*Media Literacy, Fact Checking*), la *Disinformation* e la *Malinformation* richiedono strumenti di *Threat Intelligence* e procedure di *Incident Response* simili a quelle impiegate per il crimine informatico.[4] La *Malinformation* è particolarmente insidiosa per la Governance istituzionale, poiché la difesa non può basarsi sulla smentita del contenuto, ma deve gestire la crisi di legittimità derivante dalla pubblicazione di dati reali.

1.3 L'Integrità Epistemica e il fenomeno del "Truth Decay"

Il rischio ultimo posto dal disordine informativo non è la singola notizia falsa, ma l'erosione di ciò che Lewandowsky et al. (2023) definiscono "Integrità Epistemica".[5] Le democrazie liberali e i mercati liberi funzionano sulla premessa di una "realtà condivisa" (*shared reality*): un insieme di fatti concordati su cui basare il dibattito, le *policy* e le transazioni economiche. Quando un ecosistema informativo viene inondato di rumore e manipolazione, si verifica il fenomeno del *Truth Decay* (Decadimento della Verità), analizzato approfonditamente dalla RAND Corporation.[6]

---

[4] **Wardle, C., & Derakhshan, H.** (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe Report, pp. 20-24.

[5] **Lewandowsky, S., et al.** (2023). *Misinformation and the epistemic integrity of democracy*. Current Opinion in Psychology, 54

[6] **Helmus, T. C., & Chandra, B.** (2024). *Generative Artificial Intelligence: Threats to Information Integrity*. RAND Corporation, pp. 3-5.

Il *Truth Decay* è caratterizzato da:

- un crescente disaccordo sui fatti e sulle interpretazioni analitiche di dati oggettivi;
- una sfumatura dei confini tra opinione e fatto;
- l'aumento del volume relativo e dell'influenza dell'opinione personale rispetto ai fatti;
- un calo di fiducia nelle istituzioni tradizionalmente designate come fonti di autorità fattuale (enti regolatori, media, accademia).[7]

Per le democrazie e le strutture di governo, questo si traduce in un ambiente operativo imprevedibile (*Volatile, Uncertain, Complex and Ambiguous* - VUCA). Il *Truth Decay* non è solo un fenomeno sociale, ma un meccanismo che amplifica strutturalmente il rischio sistemico. Se la legittimità delle istituzioni o la stabilità dei mercati finanziari possono essere distrutte in poche ore da una narrazione falsa che diventa virale prima di poter essere verificata, il concetto stesso di "stabilità istituzionale" viene messo in discussione.

1.4 La Velocità della Falsità e l'Infodemia

Un fattore determinante nella pericolosità del disordine informativo è la velocità di propagazione, amplificata dagli algoritmi delle piattaforme social che privilegiano l'engagement emotivo rispetto all'accuratezza. Lo studio di Vosoughi, Roy e Aral, ha analizzato la diffusione di circa 126.000 storie su Twitter (2006-2017), dimostrando empiricamente che *"il falso si diffonde significativamente più lontano, più velocemente, più in profondità e più ampiamente della verità"*. In particolare, le notizie politiche false hanno raggiunto un pubblico di 20.000 persone tre volte più velocemente rispetto alle notizie vere equivalenti. Gli autori attribuiscono questo fenomeno al fattore "novità" e alla carica emotiva (spesso disgusto o sorpresa) che le notizie false portano con sé, stimolando la ricondivisione impulsiva. Questa velocità è la ragione tecnica per cui i meccanismi di *Debunking* reattivo falliscono di fronte all'attacco.[8]

Durante crisi globali, come la pandemia COVID-19, questa dinamica ha generato una "Infodemia", ovvero una sovrabbondanza di informazioni che rende difficile per le persone trovare fonti affidabili e guide pratiche. In un contesto di crisi di sicurezza nazionale (es. un disastro naturale o un attacco terroristico), un'infodemia può paralizzare la risposta degli apparati di protezione civile o di intelligence, sommersi da "rumore" che maschera i segnali reali necessari per la gestione dell'incidente.[9]

In conclusione, il disordine informativo rappresenta una vulnerabilità strutturale dell'era digitale. La sua gestione non può essere delegata esclusivamente ai dipartimenti di

[7] **Kavanagh, J., & Rich, M. D.** (2018). *Truth Decay*. RAND Corporation, pp. 9-13.
[8] **Vosoughi, S., Roy, D., & Aral, S.** (2018). *The spread of true and false news online*. Science, 359(6380), pp. 1146–1151.
[9] **Gallotti, R., et al.** (2020). *Assessing the risks of "infodemics" in response to COVID-19 epidemics*. Nature Human Behaviour, 4(12), 1285–1293; **World Health Organization.** (2020). *Managing the COVID-19 infodemic*; e **Gallotti, R., et al.** (2020). *Assessing the risks of "infodemics"*. Nature Human Behaviour, 4, pp. 1285–1293.

comunicazione o IT, ma deve essere elevata a priorità di Governance, richiedendo un approccio integrato che unisca tecnologia, psicologia e strategia organizzativa.[10]

[10] **NIST.** (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce, p. 12.

CAPITOLO 2 - IL CONTESTO STRATEGICO: GUERRA IBRIDA E INDUSTRIALIZZAZIONE DELLA MANIPOLAZIONE

L'analisi del disordine informativo non può prescindere dal contesto geopolitico in cui si manifesta. L'uso offensivo delle informazioni non è un fenomeno isolato, ma una componente centrale delle moderne dottrine di conflitto, note come "Guerra Ibrida", e di un processo di industrializzazione della manipolazione che ha reso la disinformazione scalabile ed economica.[11]

2.1 La natura mutevole del conflitto: La Guerra Ibrida

La teoria della Guerra Ibrida ha ridefinito il concetto di sicurezza nazionale e, per estensione, della sicurezza per le infrastrutture critiche. Hoffman (2007) definisce la minaccia ibrida come la fusione simultanea e adattiva di diverse modalità di combattimento: capacità militari convenzionali, tattiche irregolari (guerriglia), atti terroristici e disordine criminale, tutti orchestrati nello stesso spazio di battaglia per ottenere obiettivi politici. In questo paradigma, la dimensione informativa non è un supporto alle operazioni cinetiche, ma un dominio operativo primario. L'obiettivo strategico spesso non è la distruzione fisica dell'avversario, ma il collasso della sua volontà politica e sociale dall'interno.[12]

Deshpande (2018) approfondisce questa dinamica evidenziando una vulnerabilità specifica delle democrazie occidentali: mentre queste tendono a concepire "guerra" e "pace" come stati binari distinti, gli attori ibridi (come Russia o Cina) operano in una "Zona Grigia" permanente. In questa zona, le azioni ostili (campagne di disinformazione massiva, attacchi *cyber* sotto la soglia di guerra dichiarata o coercizione economica) sono costanti, mantenendosi appena al di sotto del livello che scatenerebbe una risposta militare convenzionale (es. Articolo 5 della NATO).[13] Di conseguenza, il conflitto esonda dai campi di battaglia tradizionali per investire direttamente le infrastrutture civili ed economiche, rendendo la distinzione tra bersaglio militare e bersaglio civile sempre più sfumata.

[11] **Hoffman, F. G.** (2007). *Conflict in the 21st Century: The Rise of Hybrid Wars*. Potomac Institute for Policy Studies

[12] **Hoffman, F. G.** (2007). *Conflict in the 21st Century: The Rise of Hybrid Wars*. Potomac Institute for Policy Studies, p. 14.

[13] **Deshpande, V.** (2018). *Hybrid Warfare*, IDSA, pp. 5-10.

Tabella 2.1: Lo Spettro del Conflitto Ibrido e la Zona Grigia

Stato del Conflitto	Caratteristiche Operative (Cosa succede sul campo)	Ruolo dell'Informazione (La dimensione cognitiva)	Obiettivo Strategico (Il "Perché")
PACE (Stato Ideale)	**Interazione Regolata:** Diplomazia tradizionale, trattati commerciali, rispetto del diritto internazionale.	**Soft Power & Diplomazia Pubblica:** L'informazione serve a costruire attrazione culturale e fiducia. Si cerca di influenzare l'opinione pubblica estera tramite valori condivisi, non tramite l'inganno.	**Stabilità e Cooperazione:** Mantenere relazioni stabili e accrescere il prestigio nazionale attraverso la legittimità.
ZONA GRIGIA (Conflitto Ibrido)	**Aggressione Sotto-Soglia:** Cyberattacchi, coercizione economica, finanziamento di gruppi sovversivi (*proxy*), sabotaggi. Azioni ostili che rimangono *appena sotto* la soglia della risposta militare (es. Art. 5 NATO**).**	**Weaponization (Armamentizzazione):** L'informazione diventa un'arma non cinetica. Uso massiccio di disinformazione e *botnet* per polarizzare la società, confondere la realtà (*Gaslighting*) e minare la fiducia nelle istituzioni.	**Destabilizzazione Interna:** Far collassare la volontà politica e la coesione sociale dell'avversario dall'interno, senza dover combattere una guerra fisica costosa.
GUERRA CONVENZIONALE	**Scontro Cinetico:** Operazioni militari dichiarate, uso della forza fisica distruttiva (eserciti, bombardamenti), occupazione territoriale.	**Propaganda Tattica & PsyOps:** L'informazione serve a sostenere lo sforzo bellico: demoralizzare le truppe nemiche, confondere i comandi militari e mobilitare il fronte interno.	**Distruzione Fisica:** Sconfiggere militarmente l'avversario e imporre la propria volontà attraverso la superiorità di forza.

Fonte: Elaborazione basata su Hoffman (2007) e Deshpande (2018).

2.2 L'Industrializzazione della Manipolazione: Le "Lie Machines"

Se la Guerra Ibrida fornisce il "perché" dietro la necessità di minare la coesione sociale e la fiducia nelle istituzioni, l'industrializzazione della disinformazione spiega il "come" questa operazione viene condotta con una portata e un'efficacia senza precedenti. Questo fenomeno ha superato la fase artigianale e casuale per diventare un processo industriale standardizzato e scalabile, caratterizzato da sofisticate metodologie, risorse finanziarie significative e l'impiego massiccio di tecnologia avanzata.

Il Computational Propaganda Project dell'Oxford Internet Institute (OII) monitora questo fenomeno. Nei loro rapporti *Global Inventory of Organised Social Media Manipulation*, Bradshaw e Howard (2019, 2020) documentano l'ascesa delle cosiddette *"Cyber Troops"*: team organizzati da governi, forze armate o partiti politici, dedicati alla manipolazione dell'opinione pubblica *online*. Il rapporto del 2020 ha identificato la presenza di tali unità in 81 paesi, un aumento esponenziale rispetto ai 28 individuati nel 2017. Tali organizzazioni operano con budget dedicati, strutture gerarchiche e *Key Performance Indicators* (KPI), trattando la disinformazione come un servizio (*Disinformation-as-a-Service*).[14]

Figura 2.2a : Diffusione Globale delle Cyber Troops

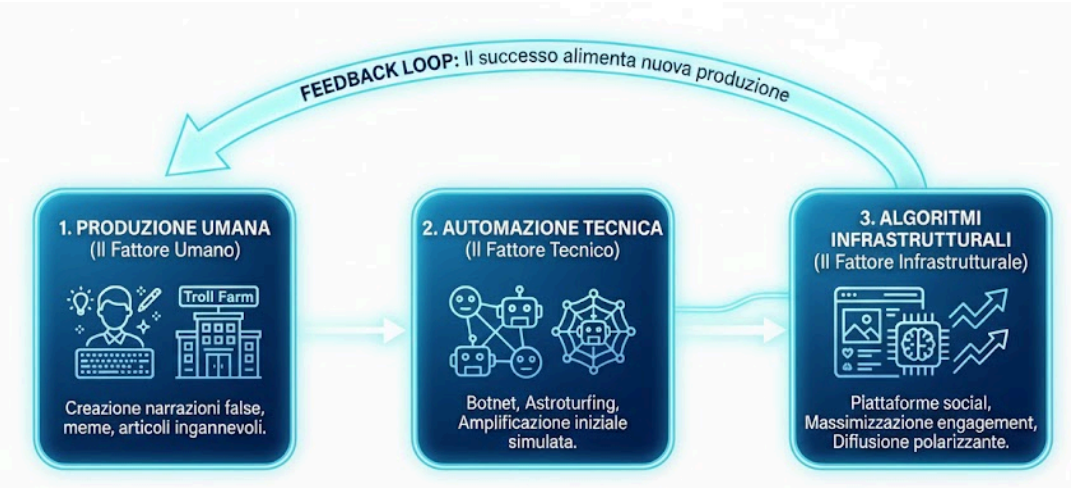
Fonte: Basata sui dati del rapporto Global Inventory of Organised Social Media Manipulation, Bradshaw e Howard (2019, 2020), (Oxford Internet Institute, 2020).

Philip Howard (2020), nel suo volume *Lie Machines*, teorizza che queste architetture di manipolazione funzionino come vere e proprie catene di montaggio, composte da tre ingranaggi interdipendenti:

[14] **Bradshaw, S., & Howard, P. N.** (2019). *The global disinformation order*; **Bradshaw, S., & Howard, P. N.** (2020). *Industrialized disinformation*. Oxford Internet Institute.

1. **Produzione di contenuti (Il fattore umano):** la creazione delle narrazioni false, dei *meme* e degli articoli ingannevoli. Spesso questa fase è esternalizzata a *"Troll Farms"* (come la nota *Internet Research Agency* di San Pietroburgo) o a società di PR oscure (aziende private mercenarie che offrono servizi di manipolazione occulta dell'opinione pubblica).
2. **Automazione (Il fattore tecnico):** l'uso di *botnet* (reti di account automatizzati) per amplificare i contenuti nelle fasi iniziali, simulando una popolarità inesistente (*astroturfing*) e ingannando i *trending topics* delle piattaforme.
3. **Algoritmi (Il fattore infrastrutturale):** lo sfruttamento degli algoritmi di raccomandazione delle piattaforme *social* (Facebook, X, TikTok), che, essendo progettati per massimizzare l'engagement, tendono a premiare e diffondere organicamente i contenuti polarizzanti e infiammatori. Il successo dell'industrializzazione dipende proprio dallo sfruttamento di questa logica di business dei veicoli di diffusione.[15]

Figura 2.2b: L'Architettura della "Lie Machine" (Macchina della Menzogna)



Fonte: Elaborazione Google Gemini basata su Howard (2020), "Lie Machines".

2.3 Il Continuum Storico: Dalle Misure Attive all'Influenza Digitale

È fondamentale sottolineare che, sebbene la tecnologia sia nuova, le dottrine sottostanti hanno radici storiche profonde. Thomas Rid (2020), nel suo studio *Active Measures*, traccia una linea diretta di continuità tra le operazioni di guerra politica condotte dal KGB sovietico durante la Guerra Fredda e le moderne campagne digitali.[16]

Le "Misure Attive" (*aktivnyye meropriyatiya*) sovietiche includevano la falsificazione di documenti, la creazione di organizzazioni di facciata e la diffusione di voci per destabilizzare le società occidentali. Come notato nei rapporti del *NATO StratCom COE* (Juurvee, 2018;

[15] **Howard, P. N.** (2020). *Lie Machines: How to Save Democracy from Troll Armies, Deceptive Ads, and Data Ops*. Yale University Press, pp. 4-8.
[16] **Rid, T.** (2020). *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux, p. 8.

NATO, 2021), i servizi di intelligence moderni hanno "resuscitato" queste tattiche adattandole all'ecosistema digitale.[17]

La rivoluzione di Internet non ha cambiato l'obiettivo (la sovversione), ma ha drammaticamente alterato i costi e i vettori di attacco:

- **Abbassamento dei costi:** un'operazione che un tempo richiedeva mesi per piantare una notizia falsa su un giornale straniero tramite agenti fisici, oggi può essere lanciata in poche ore con un costo marginale vicino allo zero. Internet ha democratizzato l'accesso alle armi di influenza.
- **Disintermediazione:** gli attaccanti possono ora bypassare i tradizionali "*gatekeeper*" dell'informazione (giornalisti, editori) per colpire direttamente i dispositivi dei cittadini e dei dipendenti *target*.[18]

2.4 Case Study Comparato: Dall'Operazione INFEKTION al COVID-19

Per comprendere la natura persistente della disinformazione, è utile comparare un'operazione analogica della Guerra Fredda con una digitale moderna.

2.4.1 Operazione INFEKTION (1983-1987)

Il KGB lanciò una massiccia campagna per diffondere la falsa teoria che il virus dell'HIV/AIDS fosse un'arma biologica creata dal Pentagono. La notizia fu piantata inizialmente su un giornale oscuro in India (*The Patriot*), per poi essere ripresa da media sovietici e infine, per "rimbalzo", da media occidentali ignari (*Trading up the chain*). Ci vollero 4 anni perché la narrazione diventasse globale.[19]

2.4.2 L'Infodemic COVID-19 (2020)

Durante la pandemia, la stessa narrazione ("Il virus è creato in laboratorio") è riemersa. Tuttavia, come analizzato da Bhattacharya & Singh (2025) , la disinformazione ha raggiunto copertura globale in *giorni*, non anni, grazie ai *social media* e alle *app* di messaggistica (WhatsApp, Telegram). La *Velocity* e la *Variety* della minaccia moderna rendono inefficaci i tempi di reazione della diplomazia tradizionale.

Conclusione per la Governance: Sebbene l'obiettivo strategico (seminare sfiducia nelle autorità sanitarie) sia rimasto identico a quello del 1983, l'accelerazione dei vettori impone alle autorità di sanità pubblica e ai governi di disporre oggi di unità di *Crisis Response inter-agenzie* capaci di operare in tempo reale (24/7).[20]

---

[17] **Juurvee, I.** (2018). *The resurrection of 'active measures'*. Hybrid CoE Strategic Analysis 7; **NATO StratCom COE.** (2021). *Social media manipulation 2021: State of the art*.

[18] **Bradshaw, S., & Howard, P. N.** (2020). *Industrialized disinformation*. Oxford Internet Institute; **Howard, P. N.** (2020). *Lie Machines*. Yale University Press.

[19] **Cull, N. J., et al.** (2017). *Soviet Subversion, Disinformation and Propaganda*. LSE Institute of Global Affairs, pp. 12-15; per il concetto di "Trading up the chain", cfr. **Marwick, A., & Lewis, R.** (2017). *Media Manipulation and Disinformation Online*. Data & Society.

[20] **Bhattacharya, S., & Singh, A.** (2025). *Unravelling the infodemic*. Frontiers in Communication; **World Health Organization.** (2020). *Managing the COVID-19 infodemic*.

CAPITOLO 3 - THREAT INTELLIGENCE: ANALISI DEGLI ATTORI E VETTORI DI ATTACCO

Una strategia di *Cyber Governance* efficace deve essere "*Threat-Informed*", guidata dalla conoscenza specifica degli avversari (*Adversary Intelligence*). Come stabilito dalla Direttiva NIS 2 (UE 2022/2555) e dalle *best practice* ISO 27001 , l'analisi del rischio deve identificare chi ha l'intenzione e la capacità di colpire l'organizzazione. Questo capitolo delinea una tassonomia operativa dei principali *Threat Actors* attivi nel dominio cognitivo, basandosi sulla reportistica di intelligence fornita da Mandiant, Google TAG e Atlantic Council.[21]

3.1 State-Sponsored Actors: La geopolitica della disinformazione

Gli attori statali rappresentano la minaccia più sofisticata (*APT - Advanced Persistent Threats*). Dispongono di risorse quasi illimitate e operano con orizzonti temporali lunghi, integrando la disinformazione nelle loro dottrine militari.

3.1.1 La Federazione Russa: Un ecosistema di propaganda

Secondo il *report* del Dipartimento di Stato USA (2022) , la Russia non si affida a un singolo canale, ma gestisce un "ecosistema di disinformazione e propaganda" composto da cinque pilastri interconnessi: comunicazioni ufficiali del governo, media finanziati dallo Stato (RT, Sputnik), fonti *proxy*, *weaponization* dei *social media* e disinformazione abilitata dal *cyber*-spionaggio.[22]

Nonostante le sanzioni imposte dal Consiglio dell'Unione Europea (2022), l'Institute for Strategic Dialogue (ISD, 2025) ha documentato la resilienza di questa rete. L'investigazione dell'ISD ha rivelato come RT continui a penetrare lo spazio informativo europeo utilizzando una rete di "siti specchio" (*mirror sites*) che replicano i contenuti per eludere i blocchi.[23]

3.1.2 La Repubblica Popolare Cinese: L'operazione "Dragonbridge"

Mentre la Russia mira spesso al caos distruttivo, la Cina persegue obiettivi di egemonia discorsiva e strategica. Mandiant (2022) e il Google *Threat Analysis Group* (TAG, 2023) hanno esposto l'operazione nota come "Dragonbridge" (o *Spamouflage*), una delle più vaste campagne di influenza multipiattaforma mai documentate. Le sue caratteristiche distintive includono il *Volume massivo* (oltre 50.000 istanze disabilitate nel 2022) e un forte *Targeting economico*. Mandiant ha documentato campagne mirate a screditare le compagnie minerarie statunitensi e australiane nel settore delle Terre Rare (*Rare Earth Elements*) , diffondendo

---

[21] **Mandiant.** (2022). *Dragonbridge: China-linked influence campaign*; **Google Threat Analysis Group.** (2023). *Over 50,000 instances of DRAGONBRIDGE activity disrupted in 2022*; **Atlantic Council.** (2020). *Iranian Digital Influence Efforts*.

[22] **United States Department of State.** (2022). *RT and Sputnik's role in Russia's disinformation and propaganda ecosystem*. Global Engagement Center.

[23] **Council of the European Union.** (2022). *Restrictive measures in response to Russia's invasion of Ukraine*; **Institute for Strategic Dialogue.** (2025). *Investigation: How Russia Today is evading sanctions in Italy*.

false notizie su presunti danni ambientali per bloccarne l'espansione. Questo è un chiaro esempio di disinformazione usata come arma di concorrenza sleale (*Unfair Competition*).[24]

3.1.3 La Repubblica Islamica dell'Iran: "Guerrilla Broadcasting"

L'Atlantic Council (2020) definisce la strategia di Teheran come "*Guerrilla Broadcasting*". Essendo tecnologicamente inferiore a USA e Israele, l'Iran adotta tattiche asimmetriche basate sull'impersonificazione, creando siti *web* che imitano testate giornalistiche legittime (es. "IUVM Press"). Tali siti aggregano notizie reali per costruire credibilità, inserendo poi strategicamente articoli di disinformazione.[25]

Tabella 3.1: Analisi Comparata dei Principali Attori Statali

Caratteristica	***Federazione Russa***	***Repubblica Popolare Cinese (Dragonbridge)***	***Iran (IUVM)***
Obiettivo Strategico	*Destabilizzazione, polarizzazione, erosione della fiducia democratica.*	*Difesa dell'immagine del PCC, attacco agli interessi economici rivali (es. Terre Rare).*	*Influenza regionale (Medio Oriente), narrativa anti-USA/Israele.*
Tattica Principale	*"Firehose of Falsehood": alto volume, canali multipli, nessuna coerenza necessaria.*	*Spamming massivo, inondazione dei canali, tentativi di mobilitazione fisica (falliti).*	*Impersonificazione di media legittimi, riciclaggio di contenuti.*
Implicazioni Economiche/ Istituzionali	*Alto (rischio sistemico, sanzioni, crisi energetiche).*	*Alto (sabotaggio di interessi nazionali chiave, spionaggio industriale).*	*Medio (rischio specifico per settore Oil & Gas e Difesa).*

Fonte: Elaborazione su dati Mandiant (2022), Google TAG (2023), Atlantic Council (2020).

[24] **Mandiant.** (2022). *Dragonbridge: China-linked influence campaign*; **Google Threat Analysis Group.** (2023). *Over 50,000 instances of DRAGONBRIDGE activity disrupted in 2022*.
[25] **Atlantic Council.** (2020). *Iranian Digital Influence Efforts: Guerrilla Broadcasting for the Twenty-First Century*. Digital Forensic Research Lab (DFRLab), pp. 4-6.

3.2 Attori Non-Statali e Terrorismo: La Guerra Cognitiva

Il conflitto moderno non è monopolio degli stati. Il *report* del DFRLab (Sadek & Mashkoor, 2023) evidenzia come i gruppi terroristici utilizzino i *social media* come "artiglieria psicologica". Due fenomeni critici emergono da questo teatro operativo:

1. **Riciclo di contenuti (Old Footage):** l'uso di video tratti da conflitti passati o persino da *videogiochi* iper-realistici (come *Arma 3*) spacciati per filmati in tempo reale.[1] Questo crea una nebbia di guerra che confonde non solo l'opinione pubblica, ma anche gli analisti di *Open-Source Intelligence* (OSINT).
2. **La narrativa dei "Crisis Actors":** una tattica difensiva insidiosa che consiste nell'accusare falsamente le vittime reali di essere attori pagati ("*Pallywood*"), disumanizzando la sofferenza e paralizzando la risposta umanitaria internazionale.[26]

3.3 Cybercrime Finanziario: L'Industrializzazione della Frode

L'integrazione dell'AI generativa nelle tattiche offensive ha reso possibile l'inganno di figure apicali e decisori critici attraverso la clonazione biometrica in tempo reale, minando la catena di comando. L'incidente che ha coinvolto la multinazionale Arup nel 2024 dimostra come la “fiducia epistemica” basata sui sensi (vista e udito) sia ormai una vulnerabilità sfruttabile.

3.3.1 Deepfake e Business Email Compromise (BEC) 2.0

La *"CEO Fraud"* si è evoluta. Nel 2019, BBC News ha riportato il caso di un'azienda energetica britannica ingannata da un *deepfake* audio del suo capo tedesco, autorizzando un bonifico urgente di 220.000 euro. L'AI ha imitato non solo la voce, ma l'intonazione e l'accento.[27]

3.3.2 Il Caso Arup: Un fallimento di Governance, non di Tecnologia

L'evento spartiacque si è verificato però nel 2024 a Hong Kong, colpendo la multinazionale ingegneristica britannica Arup. Un impiegato del settore finanziario, sospettando una truffa via *phishing*, ha richiesto una videochiamata di verifica. Gli attaccanti hanno organizzato una videoconferenza in cui il CFO (*Chief Financial Officer*) e altri colleghi erano tutti *deepfake* generati in tempo reale. Rassicurato dalla presenza visiva e vocale, l'impiegato ha autorizzato trasferimenti per 25 milioni di dollari.

Analisi del Fallimento (*Root Cause Analysis*): Questo incidente dimostra che "vedere non è più credere". Il fallimento non è stato di natura tecnologica (il *firewall* non poteva bloccare la videochiamata), ma procedurale. I protocolli di sicurezza standard non avevano previsto un protocollo di verifica *Out-of-Band* (es. una chiamata sul cellulare personale o un *token* fisico) per transazioni di tale importo, affidandosi alla validità del canale video. L'attacco cognitivo

[26] **Sadek, D., & Mashkoor, L.** (2023). *In Israel-Hamas conflict, social media become tools of propaganda and disinformation*. DFRLab Atlantic Council.
[27] **BBC News.** (2019, July 8). *Fake voice used to scam company out of €220,000*.

ha successo quando la governance istituzionale o critica fa affidamento sulla fiducia epistemica dei sensi in un canale digitale non autenticato.[28]

BOX 3.1: CASE STUDY - IL CASO ARUP E LE LEZIONI APPRESE

L'Incidente	Maggio 2024, filiale di Hong Kong della Arup.
Il Vettore	*Phishing* + *Deepfake Video/Audio* in tempo reale (GenAI).
Il Danno	Perdita finanziaria diretta di 25 milioni di USD (approx. £20m).
Lezioni per la Sicurezza Sistemica	1. ***Zero Trust* Cognitivo**: il personale critico non deve fidarsi dei propri sensi su canali digitali non autenticati. 2. ***Protocollo "Four-Eyes"***:per transazioni sopra una certa soglia, è obbligatoria l'autorizzazione congiunta. 3. ***Challenge-Response***: istruire il personale a chiedere azioni fisiche complesse (es. "girati di profilo") che i modelli attuali di *deepfake* faticano a renderizzare.[29]

3.4 Algoritmi e Diritti Umani: Il Caso "Social Atrocity" in Myanmar

L'analisi del rischio disinformativo non può limitarsi alle perdite finanziarie; deve includere i rischi legali ed etici derivanti dalla "complicità algoritmica". Il caso più emblematico di fallimento della Governance delle Piattaforme in questo ambito è documentato nel rapporto di Amnesty International (2022) intitolato *The Social Atrocity: Meta and the Right to Remedy for the Rohingya*.[30]

3.4.1 La dinamica dell'amplificazione

Nel 2017, la campagna di pulizia etnica contro la minoranza Rohingya ha visto la piattaforma Facebook non come mero spettatore passivo, ma come una cassa di risonanza attiva. Gli algoritmi di raccomandazione di Meta, progettati per massimizzare l'engagement, hanno identificato che i contenuti d'odio anti-Rohingya generavano forti reazioni emotive. Di

[28] **Hern, A.** (2024, May 17). *UK engineering firm Arup hit by deepfake scam in Hong Kong*. The Guardian.
[29] Per un'analisi delle limitazioni tecniche dei deepfake in tempo reale che rendono efficaci queste sfide fisiche, cfr. **Sensity AI.** (2020). *The State of Deepfakes: Landscape, Threats, and Impact*.
[30] **Amnesty International.** (2022). *The Social Atrocity: Meta and the Right to Remedy for the Rohingya*. Amnesty International International Secretariat.

conseguenza, il sistema ha proattivamente amplificato questi contenuti, spingendoli nei *feed* di utenti che non li avevano cercati.[31] Questo fenomeno dimostra il pericolo dei *Black Box Algorithms*: l'ottimizzazione cieca per il profitto, in assenza di *guardrail* etici, ha trasformato la piattaforma in un'arma di incitamento alla violenza.[32]

3.4.2 Il fallimento della Due Diligence

Sotto il profilo della Governance, l'indagine di Amnesty International conclude che Meta ha "contribuito in modo sostanziale" alle atrocità commesse contro i Rohingya, fallendo drammaticamente nella sua *Human Rights Due Diligence*. Il rapporto evidenzia una negligenza sistemica e prolungata: nonostante i ripetuti avvertimenti ricevuti da attivisti e organizzazioni locali fin dal 2012, l'azienda non ha allocato risorse adeguate per mitigare il rischio. Un dato emblematico citato nell'indagine rivela che, ancora nel 2014, Facebook disponeva di un solo moderatore di contenuti dedicato all'intero Myanmar, un paese di 50 milioni di abitanti nel mezzo di una crisi etnica. Questa scelta deliberata di non adattare i classificatori di odio e i team di moderazione al contesto linguistico e culturale locale stabilisce un precedente critico per tutte le multinazionali: la gestione negligente di una piattaforma digitale può portare a una complicità diretta in crimini contro l'umanità e a conseguenti richieste di risarcimento miliardarie.[33]

BOX 3.2 - OLTRE IL MODELLO COMMERCIALE: ANALISI DI ARCHITETTURE ALTERNATIVE, PROPOSTA DI SINTESI E SFIDE SISTEMICHE

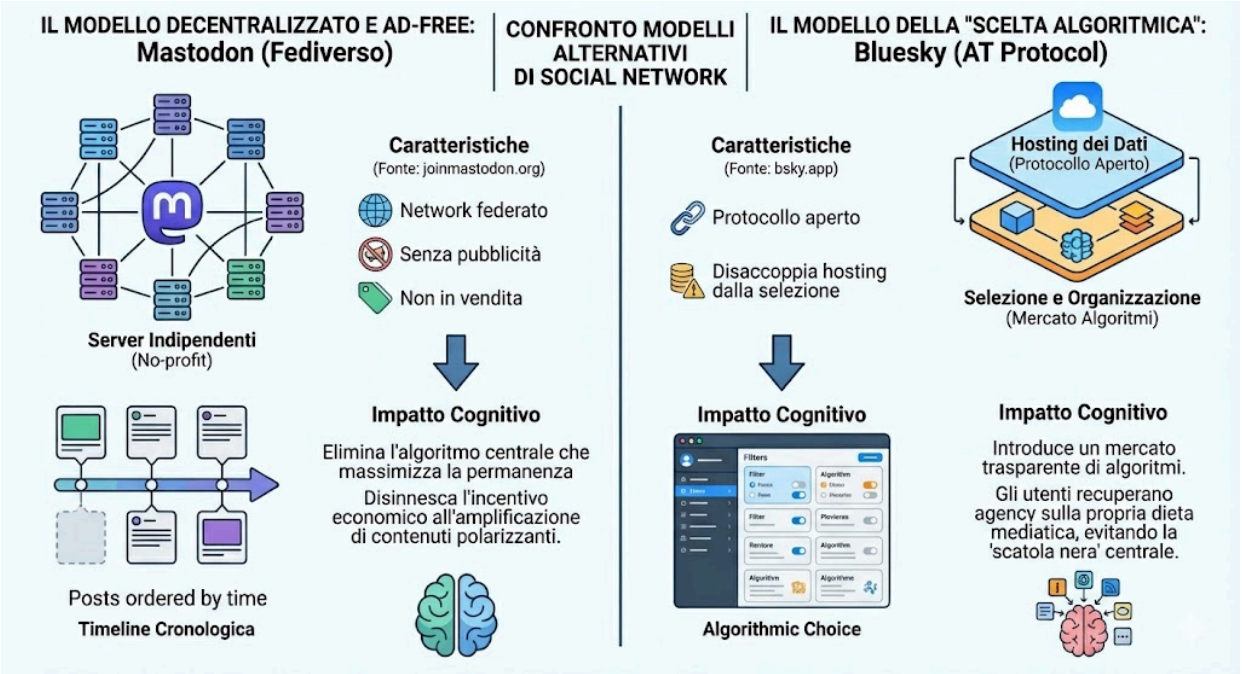
L'analisi del caso di studio sul Myanmar ha fornito l'evidenza empirica di come l'attuale modello di business delle piattaforme, basato sulla massimizzazione dell'engagement a fini pubblicitari, agisca da catalizzatore per la violenza nel mondo reale quando opera in contesti fragili. La vulnerabilità sistemica non è intrinseca alla tecnologia in sé, ma alla sua logica di business: l'algoritmo non è progettato per massimizzare la verità (*epistemic accuracy*), ma per massimizzare il tempo trascorso sulla piattaforma per vendere spazi pubblicitari.

Partendo dalla necessità urgente di rompere questo nesso strutturale tra profitto e amplificazione dell'odio, questo box esamina le architetture di social network alternativi emergenti e ne delinea una proposta teorica di sintesi per un modello resiliente. L'assunto teorico fondamentale è che rimuovendo il profitto come obiettivo primario, l'incentivo a "pompare" artificialmente i contenuti (*engagement-based ranking*) scompare, interrompendo di fatto la catena di montaggio della disinformazione analizzata nel Capitolo 2 e rendendo inefficaci tattiche come l'*astroturfing*. Infine, si esaminano in modo critico gli ostacoli al passaggio del modello concettuale a una reale infrastruttura operativa su scala mondiale.

---

[31] **Amnesty International.** (2022). *The Social Atrocity: Meta and the Right to Remedy for the Rohingya*. Amnesty International International Secretariat.
[32] **Zednik, C.** (2019). *Solving the black box problem: A normative framework for explainable artificial intelligence*. Philosophy & Technology, 33, 491–518.
[33] **Amnesty International.** (2022). *The Social Atrocity: Meta and the Right to Remedy for the Rohingya*. Amnesty International International Secretariat, pp. 35-40.

ANALISI DEI MODELLI DI RIFERIMENTO ESISTENTI



Infografica comparativa sui modelli di social network alternativi: Mastodon (decentralizzato e ad-free) e Bluesky (protocollo aperto con scelta algoritmica). Vengono evidenziate le caratteristiche principali e l'impatto cognitivo di ciascun approccio.

Se l'obiettivo di una piattaforma non è il profitto, ma il servizio pubblico (es. infrastruttura civica, accademica o statale), l'architettura può essere riprogettata dalle fondamenta. Si formula una proposta ibrida che sintetizza i punti di forza esistenti con requisiti di sicurezza proattiva:

Pilastro Architetturale Proposto	**Funzione Strategica**	**Obiettivo della Sintesi**
Struttura Non-Profit (Mutuata dal modello Federato)	**Disinnesco Economico.** La piattaforma serve l'integrità informativa come servizio pubblico, libera dalla pressione di massimizzare l'engagement per gli inserzionisti.	Garantire che le decisioni architetturali non siano mai subordinate al ROI (Ritorno sull'Investimento), eliminando la causa radice della "complicità algoritmica".

Scelta Algoritmica Trasparente (Mutuata dal modello a Protocollo)	**Mitigazione della Manipolazione.** Consente l'audit pubblico dei meccanismi di raccomandazione e offre una via d'uscita dalle *"filter bubbles"* imposte centralmente.	Evitare la creazione di un "Ministero della Verità" centrale, democratizzando e rendendo trasparente la selezione dei filtri.
Integrazione Nativa di Verifica (Requisito di Sicurezza)	**Safety by Design.** Integrazione nel *backend* di protocolli di verifica automatizzata (es. *trust scores*). Gli algoritmi devono penalizzare la visibilità di contenuti con basso punteggio di affidabilità.	Rendere la verifica un attributo infrastrutturale dell'informazione, non un'aggiunta *ex-post*. L'algoritmo sostituisce la massimizzazione dell'engagement con la massimizzazione dell'affidabilità (*trust-based ranking*).

IL REALITY CHECK: SFIDE SISTEMICHE E PROSPETTIVA STRATEGICA

Passare da questo modello teorico a una piattaforma operativa su scala globale non è solo un ostacolo ingegneristico. Rappresenta piuttosto una questione politica e geopolitica di rilevanza maggiore rispetto alle sfide affrontate dagli attuali colossi del settore.

A. La Barriera dell'Adozione e la Necessità dell'Interoperabilità: è irrealistico ipotizzare che la massa degli utenti migri verso una nuova piattaforma "sicura" solo per virtù civica, specialmente a fronte del "*lock-in*" dell'effetto rete sulle piattaforme dominanti. La resistenza cognitiva verso sistemi con meno gratificazione istantanea è elevata. L'unica via per rompere questo circolo vizioso è normativa: l'imposizione dell'interoperabilità obbligatoria ai "*walled gardens*" (ad esempio, Meta), permettendo agli utenti di migrare sulla nuova infrastruttura senza perdere la capacità di interagire con i contatti rimasti sulle vecchie piattaforme.

B. La Governance della "Verità" e il Rischio di Bias: anche con l'ausilio dell'IA, l'assegnazione di *trust scores* richiede intervento umano per il contesto culturale. Definire gli standard di "verità" rimane un problema politico irrisolto. Per evitare di replicare *bias* sistemici o creare strumenti di censura, la governance deve essere affidata a organi indipendenti, multipartitici e trasparenti (sul modello di autorità garanti o consorzi scientifici), mai direttamente all'esecutivo politico o a una singola azienda.

C. Sostenibilità Economica e il Ruolo dello Stato nella Guerra Ibrida: i costi infrastrutturali e operativi per una piattaforma globale sono proibitivi senza il modello pubblicitario. Tuttavia, il contesto di Guerra Ibrida permanente cambia l'equazione. Se lo spazio informativo è un dominio di conflitto, un'infrastruttura social resiliente alla manipolazione ostile diventa un asset di sicurezza nazionale, analogo a una rete di difesa critica. Ciò giustifica un finanziamento pubblico strategico (es. consorzi internazionali di democrazie), a patto che sia blindato dalle garanzie di indipendenza (punto B) per scongiurare derive autoritarie.

Questa proposta non descrive una semplice innovazione di prodotto, ma il prototipo di una infrastruttura pubblica digitale strategica. La sua realizzazione richiede un nuovo patto sociale e politico che riconosca la sicurezza cognitiva dei cittadini come un bene pubblico fondamentale da finanziare e proteggere, affrontando il difficile equilibrio tra la necessità di difesa statale e le garanzie irrinunciabili di libertà di espressione.[34]

---

[34] La proposta delineata in questo riquadro è il frutto di una riflessione congiunta. Lo sviluppo del framework concettuale è stato successivamente sottoposto a un confronto dialettico con diverse intelligenze artificiali (Gemini e ChatGPT), utilizzate come strumento di supporto per vagliarne la fattibilità tecnica, mapparne i limiti critici e stimolare ulteriori prospettive di indagine poi integrate nel testo definitivo.

CAPITOLO 4 - IL VETTORE TECNOLOGICO: INTELLIGENZA ARTIFICIALE E AUTOMAZIONE

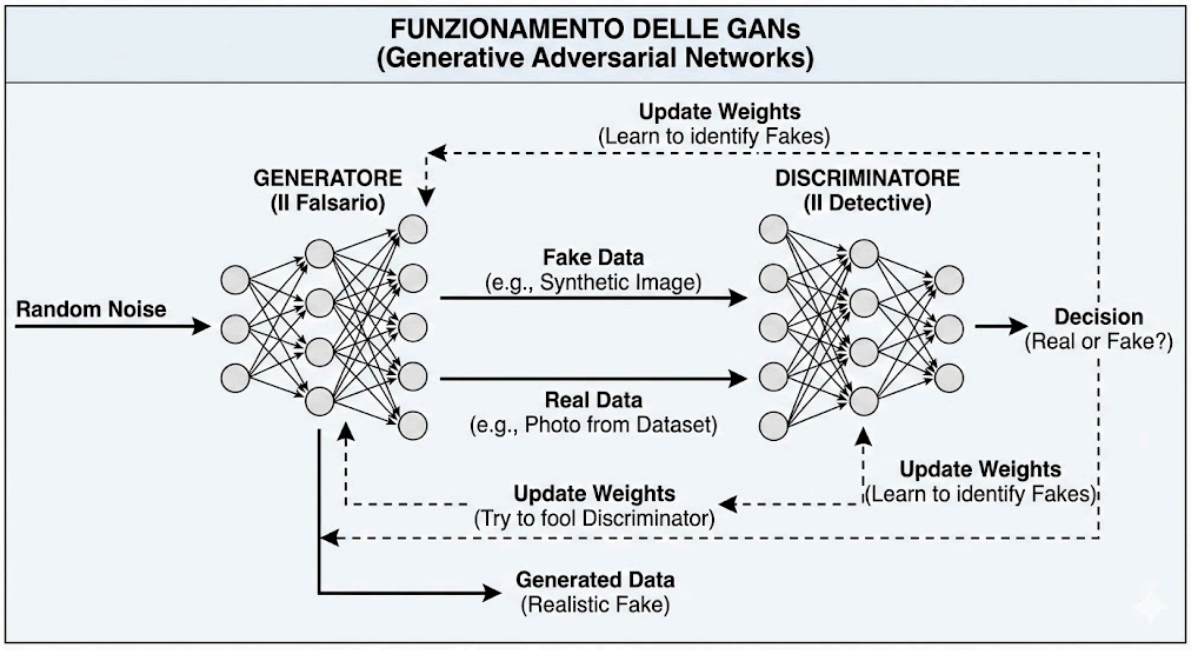
L'efficacia della disinformazione moderna risiede nella capacità tecnologica di scalare l'inganno. L'era digitale ha rimosso il vincolo umano della produzione di contenuti attraverso l'automazione e l'Intelligenza Artificiale Generativa (GenAI).[35]

4.1 Il Motore della Falsificazione: Generative Adversarial Networks (GANs)

La tecnologia abilitante alla base dei moderni *Synthetic Media* (deepfake) sono le *Generative Adversarial Networks* (GANs), introdotte nel 2014 dai ricercatori dell'Università di Montréal guidati da Ian Goodfellow. Questa architettura si basa su un gioco a somma zero tra due reti neurali distinte che competono tra loro:

1. **Il Generatore (The Forger):** ha il compito di creare un dato sintetico (es. un volto umano) partendo da rumore casuale, cercando di ingannare il discriminatore.
2. **Il Discriminatore (The Detective):** ha il compito di analizzare il dato e classificare come autentico (proveniente dal *dataset* di *training*) o sintetico (creato dal generatore).

Figura 4.1 : Architettura e Processo di Addestramento delle Generative Adversarial Networks (GANs)



Fonte: Elaborazione grafica Google Gemini basata sull'architettura originale descritta in ***Goodfellow, I., et al.*** *(2014).*

[35] **Howard, P. N.** (2020). *Lie Machines: How to Save Democracy from Troll Armies, Deceptive Ads, and Data Ops*. Yale University Press, pp. 1-15; cfr. **Bradshaw, S., & Howard, P. N.** (2020). *Industrialized disinformation: 2020 global inventory of organized social media manipulation*. Oxford Internet Institute.

Attraverso milioni di cicli di addestramento (*Backpropagation*), il Generatore impara a produrre falsi così accurati che il Discriminatore non riesce più a distinguerli dalla realtà.[36] Il report Sensity AI (2020), *The State of Deepfakes*, evidenzia che questa tecnologia è passata da curiosità accademica a strumento accessibile, con un'impennata nell'uso per spionaggio aziendale e frodi biometriche dal 2023. L'architettura GANs illustra tecnologicamente l'asimmetria attacco/difesa, dove la sola *detection* è strutturalmente destinata a inseguire la generazione.[37]

4.2 L'Amplificatore: Social Bots e "Cyborg"

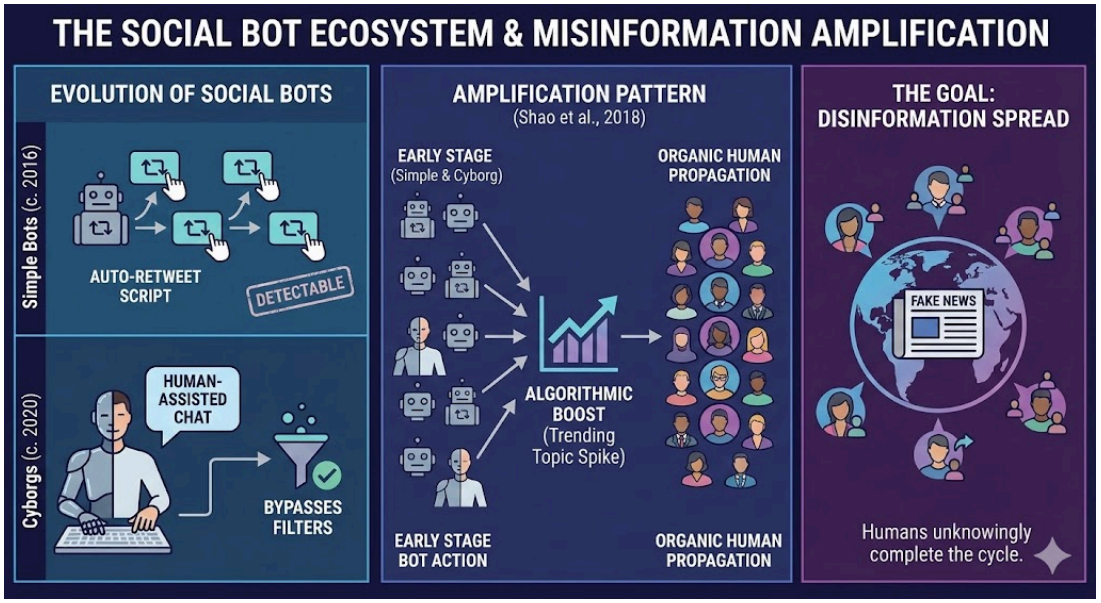
Se le GAN creano il contenuto, i *Bot* ne garantiscono la distribuzione capillare. Ferrara et al. (2016), definiscono i *bot* sociali come algoritmi informatici che controllano automaticamente *account* sui *social media* per imitare il comportamento umano.[38] Tuttavia, la minaccia si è evoluta. Come notato da Cresci (2020), i *bot* moderni non sono più semplici script che retwittano a raffica (facilmente rilevabili), ma sono diventati *"Cyborg"*: account ibridi gestiti in parte da *software* e in parte da umani, capaci di sostenere conversazioni coerenti e aggirare i filtri anti-spam.[39]

Shao et al. (2018), analizzando la diffusione di 400.000 articoli su Twitter, hanno scoperto un *pattern* operativo preciso: i *bot* agiscono nelle primissime fasi della vita di una notizia falsa. Il loro ruolo è colpire l'algoritmo di raccomandazione della piattaforma entro pochi secondi dalla pubblicazione, generando un picco artificiale di *engagement* che spinge la notizia nei *Trending Topics*. Da quel momento, gli utenti umani reali, percependo la notizia in tendenza, la condividono organicamente, completando inconsapevolmente l'opera di disinformazione. La botnet funge così da innesco algoritmico, trasformando l'attacco da tecnico a psicologico.[40]

---

[36] **Goodfellow, I., et al.** (2014). *Generative adversarial nets*. In *Advances in neural information processing systems* (pp. 2672-2680).
[37] **Sensity AI.** (2020). *The State of Deepfakes: Landscape, Threats, and Impact*.
[38] **Ferrara, E., et al.** (2016). *The rise of social bots*. Communications of the ACM, 59(7), 96-104.
[39] **Cresci, S.** (2020). *A decade of social bot detection*. Communications of the ACM, 63(10), 72-83.
[40] **Shao, C., et al.** (2018). *The spread of low-credibility content by social bots*. Nature communications, 9(1), 4787.

Figura 4.2: Evoluzione dei Social Bots e Meccanismo di Amplificazione Algoritmica della Disinformazione



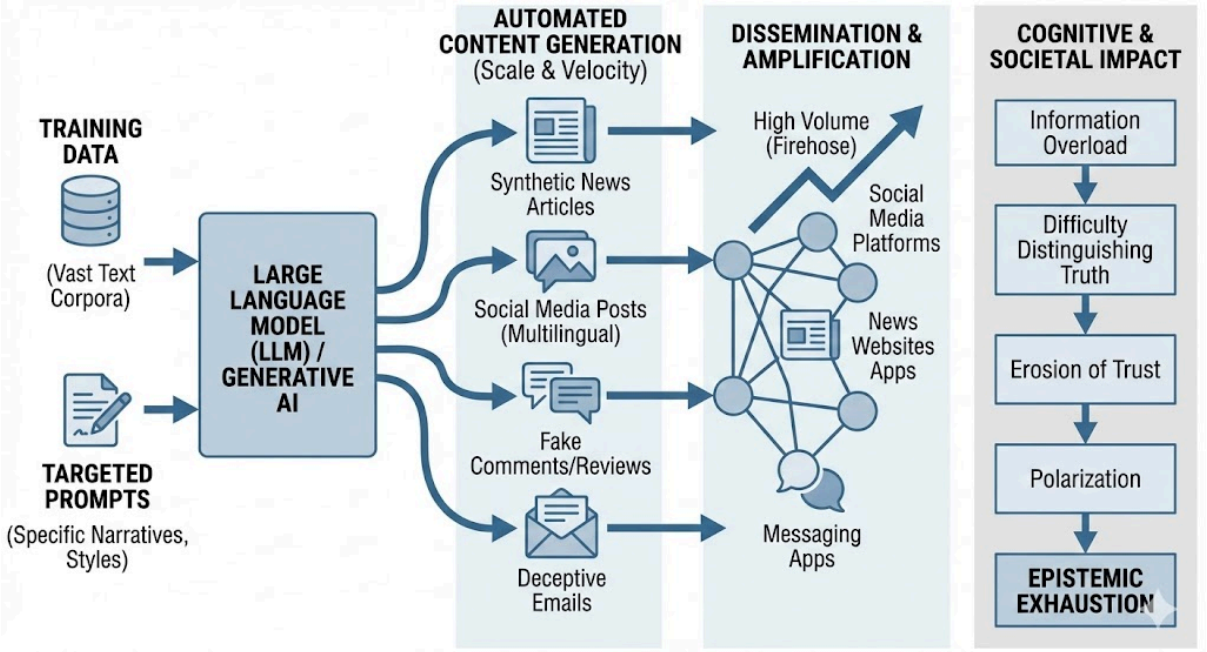
Fonte: Elaborazione grafica Google Gemini basata sui concetti e i dati descritti in ***Ferrara, E., et al.*** *(2016).;* ***Cresci, S.*** *(2020). ;* ***Shao, C., et al.*** *(2018).*

4.3 L'Automazione dell'Inganno: Il ruolo dei Large Language Models (LLM)

L'avvento dei *Large Language Models* (LLM) e dell'Intelligenza Artificiale Generativa segna un punto di flesso (*inflection point*) nella storia delle operazioni di influenza. Esiste un ampio consenso scientifico sul fatto che questa tecnologia abbia rimosso le storiche barriere logistiche che limitavano la scala della propaganda, in particolare il "collo di bottiglia" umano di produzione di contenuti credibili in diverse lingue. Oggi, i modelli generativi possono produrre volumi illimitati di testi persuasivi in dozzine di lingue simultaneamente, con una padronanza stilistica e grammaticale indistinguibile da quella nativa. Questa perfezione formale rende obsoleti gli indicatori euristici (come i tipici errori grammaticali nelle truffe online del passato) che permettevano agli utenti di riconoscere facilmente i tentativi di inganno, conferendo una credibilità immediata anche alle narrazioni più artefatte.[41]

[41] **Helmus, T. C., & Chandra, A.** (2024). *Artificial intelligence and disinformation: The new frontier of influence operations*. RAND Corporation; e il rapporto fondativo di **Goldstein, J. A., et al.** (2023). *Generative AI for Information Operations: A Framework for Threat Assessment*. Center for Security and Emerging Technology (CSET), OpenAI, and Stanford Internet Observatory.

Figura 4.3: L'Impatto dei Large Language Models sull'Automazione della Disinformazione e la Sicurezza Cognitiva



Fonte: Elaborazione grafica Google Gemini basata sui framework di analisi delle minacce e sugli impatti cognitivi descritti in ***Goldstein, J. A., et al.*** *(2023);* ***Helmus, T. C., & Chandra, A.*** *(2024).*

L'implicazione strategica più allarmante è l'accelerazione esponenziale della tattica nota come "*Firehose of Falsehood*" (Idrante di menzogne). Nell'era dell'IA generativa, si crea un'asimmetria strutturale ed economica insostenibile tra attacco e difesa: mentre un attore malevolo può generare migliaia di variazioni di una notizia falsa in pochi secondi a un costo marginale vicino allo zero, il processo di verifica (*fact-checking*) da parte di giornalisti e ricercatori rimane un'attività lenta e costosa. L'obiettivo finale di questo bombardamento informativo automatizzato trascende la semplice persuasione su un singolo tema. Mira piuttosto a generare un *esaurimento epistemico* nella popolazione target. Sommersi da un rumore di fondo costante dove l'autentico è sempre più difficile da distinguere dal falso generato dalla macchina, i cittadini tendono a perdere fiducia in qualsiasi fonte istituzionale e a ritirarsi dal dibattito pubblico, accelerando la frammentazione sociale desiderata dagli attaccanti. Il costo marginale quasi nullo per l'attaccante si traduce in un costo cognitivo insostenibile per il ricevente.[42]

[42] Il concetto di asimmetria dei costi è centrale in **Goldstein, J. A., et al.** (2023), pp. 15-18. L'impatto cognitivo di saturazione ed esaurimento è discusso in **Helmus, T. C., & Chandra, A.** (2024). Per il contesto storico dell'industrializzazione della disinformazione, cfr. anche **Howard, P. N.** (2020). *Lie Machines*.

CAPITOLO 5 - LA VULNERABILITÀ CRITICA: IL FATTORE UMANO E I *BIAS* COGNITIVI

Se la tecnologia (AI, Bot) fornisce il vettore di attacco, la mente umana rimane la vulnerabilità critica sfruttata dagli attaccanti. In un framework di Cyber Governance, i dipendenti e gli *stakeholder* non sono semplici utenti, ma "nodi della rete" suscettibili a compromissione non tramite *malware*, ma tramite manipolazione psicologica.[43]

5.1 Oltre la Partigianeria: Il Fenomeno del "Lazy Reasoning"

Un errore comune nella gestione del rischio disinformativo è assumere che le persone credano alle *fake news* solo perché confermano le loro convinzioni politiche preesistenti (*Motivated Reasoning*). Tuttavia, lo studio di Pennycook & Rand (2019), intitolato *Lazy, not biased*, dimostra empiricamente il contrario. Attraverso una serie di esperimenti comportamentali, gli autori hanno rilevato che la suscettibilità alle *fake news* è correlata primariamente alla mancanza di pensiero analitico riflessivo (*Lazy Reasoning*), piuttosto che al *bias* partigiano.[44]

In un contesto operativo critico, anche personale altamente qualificato e politicamente neutro può cadere vittima di una campagna di disinformazione se si trova in condizioni di stress o sovraccarico cognitivo. Gli attaccanti cercano di indurre una reazione rapida e poco ponderata (sfruttando il *Sistema 1*, il pensiero intuitivo e automatico descritto da Daniel Kahneman) per evitare che le vittime si fermino a valutare criticamente la situazione (attivando il *Sistema 2*, il pensiero logico e analitico).[45]

5.2 Meccanismi di Persistenza: Perché il Debunking Fallisce

Perché è così difficile correggere una falsa credenza una volta che si è radicata? Lewandowsky et al. (2017), identificano due bias cognitivi che rendono inefficaci le smentite tradizionali (*Debunking*):

1. ***Illusory Truth Effect*** **(Effetto Verità Illusoria):** la semplice ripetizione di un'affermazione, anche se palesemente falsa, aumenta la probabilità che venga percepita come vera. La saturazione dei canali crea una "familiarità" con la menzogna che il cervello confonde con la "verità".
2. ***Continued Influence Effect*** **(Effetto dell'Influenza Continua):** gli studi mostrano che, anche dopo che una notizia è stata ufficialmente smentita e la persona ha

[43] **Lewandowsky, S., Ecker, U. K. H., & Cook, J.** (2017). *Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era*. Journal of Applied Research in Memory and Cognition, 6(4), pp. 353–369; **Pennycook, G., & Rand, D. G.** (2019). *Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning*. Cognition, 188, pp. 39–50.

[44] **Pennycook, G., & Rand, D. G.** (2019). *Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning*. Cognition, 188, pp. 39–50.

[45] Il framework duale "Sistema 1 / Sistema 2" è teorizzato nell'opera fondamentale di economia comportamentale e psicologia cognitiva: **Kahneman, D.** (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

accettato la smentita, l'informazione falsa continua a influenzare i suoi ragionamenti futuri.[46]

La smentita lascia un *vuoto causale* nella narrazione mentale dell'evento. Se la realtà non fornisce una spiegazione alternativa completa (una contro-narrazione), il cervello tornerà a usare la spiegazione falsa in quanto ritenuta "meglio di niente". L'inefficacia del *Debunking* è la giustificazione scientifica per un approccio proattivo (*Prebunking*).

Tabella 5.2: Tassonomia dei Principali Bias Cognitivi Sfruttati nella Disinformazione

Bias Cognitivo	**Descrizione sintetica**	**Effetto sulla percezione**	**Esempio tipico**
Bias di conferma (confirmation bias)	Tendenza a cercare e credere solo a informazioni coerenti con le proprie convinzioni pregresse.	Rafforza la polarizzazione e riduce l'apertura a dati contrari.	Gli utenti condividono notizie false che confermano la propria posizione politica.
Effetto verità illusoria (illusory truth effect)	Ripetizione costante di una falsità aumenta la sensazione di veridicità.	La familiarità diventa criterio di credibilità.	Teorie cospirazioniste rilanciate ciclicamente sui social diventano "verità percepite".
Euristica della disponibilità (availability heuristic)	Le persone giudicano più probabile ciò che ricordano facilmente o che suscita emozione.	Amplifica la percezione del rischio e dell'urgenza.	Notizie di crimini isolati fanno credere a un aumento generalizzato della violenza.
Effetto di ritorno (backfire effect)	La correzione di una notizia falsa può rafforzare la convinzione errata.	Genera resistenza alle smentite e sfiducia nei media ufficiali.	Fact-checking sui vaccini percepito come censura "di sistema".
Bias di gruppo (ingroup bias)	Propensione a considerare attendibili solo le fonti interne al proprio gruppo identitario.	Rafforza la tribalizzazione digitale e la diffusione interna di fake news.	Comunità online che condividono solo contenuti prodotti "dai loro".

[46] **Lewandowsky, S., Ecker, U. K. H., & Cook, J.** (2017). *Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era*. Journal of Applied Research in Memory and Cognition, 6(4), pp. 353–369.

Effetto bandwagon (bandwagon effect)	Tendenza a credere o aderire a un'idea perché è percepita come popolare.	Favorisce la viralità e la legittimazione sociale della falsità.	Un hashtag virale spinge utenti neutrali a credervi per conformismo.

*Fonte: Elaborazione basata sulla revisione della letteratura di psicologia cognitiva e comportamentale (**Kahneman**, 2011; **Lewandowsky et al.**, 2017).*

5.3 Identità Sociale e Vulnerabilità Democratica

La disinformazione moderna ha superato la logica "*broadcasting*" (un messaggio unico per tutti) per abbracciare un approccio *Targeted* e chirurgico. Elsa Hedling (2025), analizza come gli attori ibridi sfruttino i *Big Data* per profilare i bersagli non solo in base ai loro interessi (come nel marketing tradizionale), ma specificamente in base alla loro identità sociale di gruppo (religiosa, politica, professionale, etnica).[47]

Il meccanismo è insidioso: l'attaccante confeziona un contenuto progettato specificamente per essere percepito come una minaccia esistenziale a quell'identità (es. "La politica aziendale X sta umiliando i lavoratori della tua categoria", oppure "L'azienda Y finanzia segretamente gruppi che odiano i tuoi valori"). Di fronte a tale minaccia percepita, non si attiva una risposta razionale, bensì una reazione viscerale nota come *Affective Polarization* (Polarizzazione Affettiva). In questo stato psicologico di "noi contro loro", la priorità cognitiva dell'individuo cessa di essere la ricerca della verità fattuale (*epistemic accuracy*); l'unico obiettivo diventa la difesa del proprio gruppo ("*ingroup*") e l'ostilità verso il gruppo avverso percepito ("*outgroup*").

Per le istituzioni democratiche, le implicazioni di questa dinamica sono sistemiche e potenzialmente paralizzanti. Gli attacchi informativi più pericolosi non mirano più a diffondere una singola menzogna su un candidato, ma a *weaponizzare* le fratture ideologiche, etniche o sociali preesistenti (le cosiddette *culture wars*). Un attore statale ostile può orchestrare campagne mirate per mobilitare segmenti della popolazione contro le istituzioni (es. magistratura, organi di stampa, forze dell'ordine) o per polarizzare il dibattito pubblico su temi etici sensibili (es. migrazione, diritti civili). In uno stato di polarizzazione affettiva intensa, i cittadini smettono di cercare la verità fattuale e diventano impermeabili ai dati oggettivi forniti dalle autorità o dai media tradizionali, percependo l'avversario politico non come un competitor, ma come una minaccia esistenziale o un "nemico morale" da sconfiggere, rendendo impossibile la formazione di un consenso su cui basare le *policy*.[48]

---

[47] Per l'applicazione di questo concetto alle minacce ibride, cfr. **Hedling, E.** (2025). *Social identities and democratic vulnerabilities*. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE), Paper 24. Il concetto teorico di "Polarizzazione Affettiva" (la tendenza a nutrire antipatia e sfiducia verso i membri dell'outgroup politico/sociale, indipendentemente dalle loro posizioni ideologiche) è definito nell'opera seminale di **Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J.** (2019). *The origins and consequences of affective polarization in the United States*. Annual Review of Political Science, 22, pp. 129-146.

[48] **Hedling, E.** (2025). *Social identities and democratic vulnerabilities: Learning from examples of targeted disinformation*. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE), Paper 24.

CAPITOLO 6 - STRATEGIE DI DIFESA E GOVERNANCE: DALLE NORME ALLA RESILIENZA ATTIVA

La fase propositiva richiede un superamento della mera ricognizione delle soluzioni già presenti in letteratura e nelle prassi correnti. Tali soluzioni sono state considerate, ma sottoposte a verifica per valutarne l'effettiva adeguatezza rispetto alla minaccia delineata nei capitoli precedenti, con l'obiettivo di individuare anche eventuali alternative più resilienti.

L'analisi è stata condotta a partire dai contenuti del corso *Governance della Cybersecurity* e dalle vulnerabilità emerse nei capitoli precedenti. Su questa base, il problema è stato discusso con diversi modelli di IA generativa (ChatGPT e Gemini), utilizzati per esplorare possibili configurazioni, mettere in evidenza criticità e valutare la tenuta delle soluzioni esistenti. Le ipotesi emerse sono state successivamente filtrate e trattenute solo quando coerenti e sostenute da una logica operativa solida.

Il lavoro si è articolato secondo la seguente pipeline:

1. **Definizione del problema**, fondata sul quadro concettuale del corso e sull'analisi delle vulnerabilità.

2. **Confronto strutturato** tra soluzioni documentate e ipotesi discusse con i modelli LLM.

3. **Selezione delle opzioni maggiormente consistenti**, privilegiando quelle che mostravano robustezza teorica e potenziale applicabilità.

Il Framework presentato in questo capitolo costituisce pertanto l'esito di questo processo: una sintesi delle configurazioni che, nel confronto tra teoria, prassi ed esplorazione simulata, hanno mostrato la maggiore solidità.

6.1 Il Framework Normativo: Regolazione e Gestione del Rischio Sistemico

Il panorama normativo europeo offre oggi strumenti potenti per strutturare la difesa, con il *Digital Services Act* (DSA) che si configura come il *gold standard* regolatorio. Come analizzato da Ó Fathaigh, Buijs e van Hoboken (2025), gli Articoli 34 e 35 del DSA segnano un passaggio fondamentale: impongono alle grandi piattaforme (*VLOPs*) un preciso obbligo di gestione del rischio sistemico. Tuttavia, emerge una chiara distinzione tra il piano della realtà normativa e quello della necessità operativa. Mentre il DSA definisce rigorosamente il perimetro dell'obbligo legale (il *"cosa"*), esso non prescrive uno standard tecnico specifico per la misurazione del rischio algoritmico (il *"come"*). In assenza di uno standard armonizzato europeo, il presente elaborato individua nel NIST AI Risk Management Framework lo strumento metodologico più maturo per colmare questa lacuna operativa.[49]

[49] **Ó Fathaigh, R., Buijs, D., & van Hoboken, J.** (2025). *The Regulation of Disinformation Under the Digital Services Act*. Media and Communication, 13.

Pertanto, la strategia di *governance* proposta si fonda sull'integrazione necessaria di questi due livelli:

1. **Controllo dell'Integrità (Livello Normativo - DSA):** Per ottemperare agli obblighi di gestione del rischio sistemico (artt. 34 e 35), si impone alle piattaforme un dovere di *due diligence* sull'affidabilità dei flussi informativi. Ciò comporta l'obbligo giuridico di: 1) valutare e mitigare i rischi derivanti da nodi di diffusione inaffidabili; 2) fornire trasparenza sui parametri dei sistemi algoritmici (art. 27 DSA).
2. **Operativizzazione Tecnica (Livello Metodologico - NIST):** L'adozione delle fasi operative del NIST AI RMF (*Map, Measure, Manage*) come protocollo standard per mappare i rischi dell'IA generativa e produrre le evidenze di conformità richieste dal regolatore.[50]

6.1.1 Il "Gap Operativo": L'AI Act, la carenza di standard armonizzati e la necessità di un ponte metodologico transitorio

Sebbene l'Unione Europea si sia dotata del primo quadro normativo completo sull'Intelligenza Artificiale con l'*AI Act* (Regolamento UE 2024/1689), l'applicazione pratica di tali norme alla sicurezza cognitiva si scontra oggi con un "gap operativo" critico. L'Articolo 9 del Regolamento impone l'adozione di un sistema di gestione dei rischi (*Risk Management System*), ma gli standard tecnici armonizzati necessari per rendere operativi questi obblighi sono ancora in fase di elaborazione e non saranno pienamente disponibili prima della fine del 2025. In questo scenario di *vacatio* tecnica, le organizzazioni si trovano di fronte al paradosso di dover ottemperare agli obblighi di *due diligence* immediati previsti dal *Digital Services Act* (DSA) – in particolare la mitigazione dei rischi sistemici ex artt. 34 e 35 – senza disporre di una metrica europea standardizzata per quantificare il rischio algoritmico. Come evidenziato dal rapporto OECD *"Facts not Fakes"* (2024), sebbene il DSA richieda audit indipendenti, manca ancora una definizione univoca di *"standard specifici e quantificabili"* che rendano tali valutazioni comparabili ed efficaci.[51]

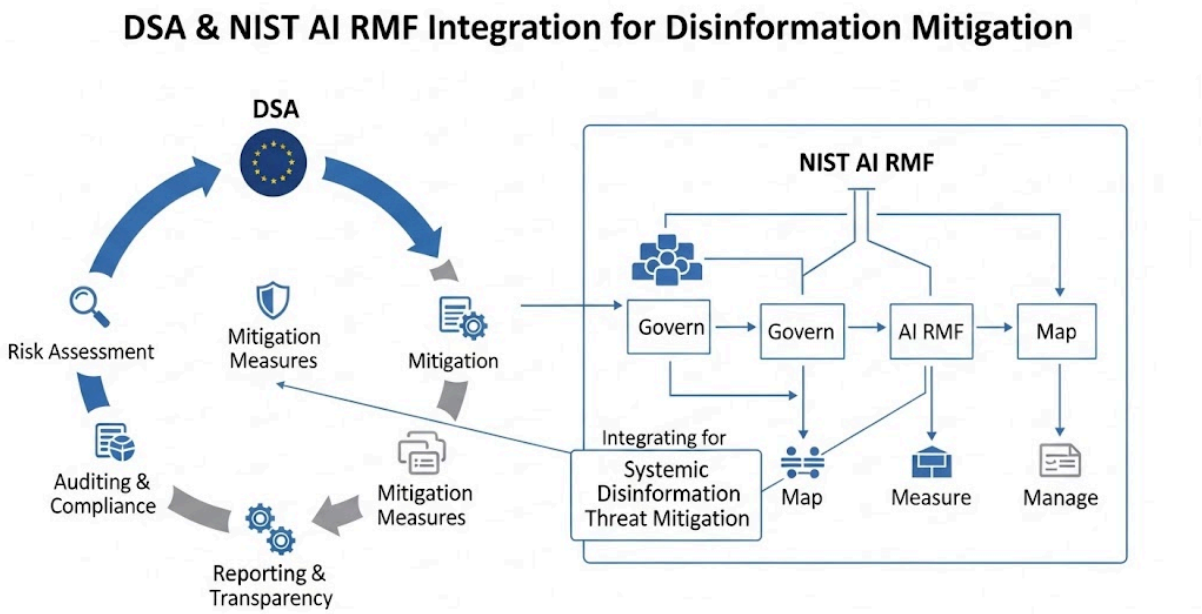
La letteratura scientifica più recente conferma questa frammentazione. Marushchak et al. (2025), analizzando le strategie di contrasto alla disinformazione potenziata dall'IA, trattano le normative europee (DSA, AI Act) e gli standard tecnici statunitensi come ambiti distinti e geograficamente separati. Tuttavia, gli stessi autori riconoscono che il NIST AI Risk Management Framework (AI RMF), sviluppato negli Stati Uniti, rappresenta oggi lo standard

[50] La proposta di integrazione tra il framework normativo europeo (DSA) e lo standard tecnico statunitense (NIST) è frutto di un'elaborazione supportata dall'uso di sistemi di Intelligenza Artificiale Generativa. L'IA è stata utilizzata per analizzare i materiali del corso, identificando nella mancanza di uno standard tecnico operativo la principale vulnerabilità dell'attuale impianto normativo e suggerendo il NIST AI RMF come soluzione metodologica coerente con i principi di governance appresi.
NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, U.S. Department of Commerce.
[51] **OECD** (2024), *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity*, OECD Publishing, Paris.

più maturo per identificare specificamente i rischi di provenienza e manipolazione del contenuto generativo (Deepfakes, Disinformazione).[52]

La presente proposta di governance intende superare questa dicotomia geopolitica attraverso una strategia di "interoperabilità pragmatica". In attesa del completamento degli standard armonizzati UE, si individua nel NIST AI RMF il "ponte tecnico" ideale per operazionalizzare gli obblighi del DSA. L'adozione del framework NIST non costituisce una sottomissione a standard extra-UE, bensì una misura transitoria di *best practice* che risponde all'urgenza della minaccia ibrida attuale, fornendo alle aziende e ai regolatori una metodologia ingegneristica (Map, Measure, Manage) immediatamente applicabile per soddisfare i requisiti di legge europei sulla sicurezza cognitiva.

Figura 6.1: L'Integrazione dei Framework Regolatori (DSA e NIST AI RMF) per la Governance del Rischio Sistemico



Fonte: Elaborazione grafica Google Gemini basata sull'analisi di ***Ó Fathaigh, R., et al.*** *(2025) e* ***NIST.*** *(2023).*

6.2 Verifica della Resilienza: Adversarial Testing

Non si può difendere ciò che non si è testato. Staves et al. dimostrano l'efficacia dell'*Adversary-Centric Security Testing* (simulazione di attacco) negli ambienti operativi e di infrastruttura critica, un approccio che deve essere esteso al dominio informativo strategico. La simulazione deve testare la capacità dei decisori di resistere sotto stress cognitivo (*Lazy Reasoning* in azione).

Gli attori statali e le istituzioni preposte alla sicurezza dovrebbero condurre periodicamente :

[52] **Marushchak A., Petrov S., Khoperiya A.** (2025), *Countering AI-powered disinformation through national regulation: learning from the case of Ukraine*, Frontiers in Artificial Intelligence, Vol. 7.

- ***Disinformation Tabletop Exercises****:*simulazioni in cui i decisori strategici governativi o di apparati di sicurezza devono reagire a una crisi informativa in tempo reale, con *input* falsi generati da AI.
- ***Test di Phishing Cognitivo****:* sottoporre il personale critico non solo a *phishing* tecnico, ma a narrazioni false verosimili che minano la coesione o la fiducia interna, al fine di misurare la prontezza di risposta e l'adesione all'inganno senza verifica. Il *Phishing Cognitivo* è l'equivalente pratico di un attacco LLM mirato e serve a validare l'efficacia dei protocolli *Zero Trust* applicati alla sfera cognitiva.[53]

6.3 Difesa Cognitiva Attiva: Prebunking e Inoculazione Psicologica

La difesa più promettente emersa dalla ricerca recente è il *Prebunking* (o Inoculazione). Roozenbeek & van der Linden (2019) hanno dimostrato scientificamente l'efficacia di questo approccio. Lo studio prova che esporre preventivamente le persone a versioni "indebolite" delle tecniche di manipolazione (es. spiegare come funziona un *bot* o l'uso del linguaggio emotivo per manipolare) conferisce una resistenza immunitaria cognitiva.

Invece di aspettare che una *fake news* si diffonda per poi smentirla (reazione), le istituzioni devono educare preventivamente i cittadini e il personale critico sulle tattiche che verranno usate contro di loro (prevenzione). In conclusione, la *Cyber Governance* moderna richiede di spostare il *budget* dalla sola protezione perimetrale alla costruzione di una *Resilienza Cognitiva*: una popolazione capace di pensiero critico e processi decisionali blindati contro l'inquinamento informativo. [54]

[53] **Staves, A., Gouglidis, A., & Hutchison, D.** (2020). *An Analysis of Adversary-Centric Security Testing within Information and Operational Technology Environments*. Lancaster University

[54] **Roozenbeek, J., & van der Linden, S.** (2019). *Fake news game confers psychological resistance against online misinformation*. Palgrave Communications, 5(65).

CONCLUSIONI

L'analisi svolta evidenzia come la manipolazione informativa abilitata dalla *GenAI* si configuri oggi come una minaccia strutturale per la stabilità delle società aperte. L'automazione della produzione di contenuti persuasivi, unita alla capacità degli attori ostili di operare in modo continuo e sotto-soglia, produce un vantaggio strategico difficilmente colmabile attraverso approcci tradizionali. La radice di questa asimmetria non è solo tecnologica: risiede nella vulnerabilità cognitiva dei decisori e nella fragilità dei processi collettivi di formazione della verità.

Il lavoro ha mostrato che la difesa contro tali dinamiche non può basarsi esclusivamente su strumenti di carattere tecnico né su misure reattive. La sfida riguarda la protezione dell'integrità epistemica, la gestione del rischio informativo e la capacità delle istituzioni di mantenere processi decisionali resilienti in un contesto caratterizzato da saturazione, polarizzazione e inganno automatizzato.

In questo quadro, è stato proposto un modello di governance fondato sull'integrazione tra leve normative, metodologie di valutazione del rischio e strumenti di difesa attiva. Tale modello non rappresenta una soluzione definitiva, ma un approccio orientato a ridurre la frammentazione e a costruire un coordinamento più solido tra dimensione regolatoria, tecnica e cognitiva. Il valore dell'approccio è rafforzato dall'adozione di una metodologia esplorativa che ha combinato fonti consolidate con un processo di simulazione dialogica tramite modelli generativi, utile a testare ipotesi, mettere in evidenza criticità e valutare la resilienza delle diverse opzioni.

Restano tuttavia aperte questioni centrali. Tra queste: la definizione di metriche condivise per misurare l'impatto cognitivo delle manipolazioni, la necessità di standard tecnici armonizzati per la gestione del rischio algoritmico e il problema del bilanciamento tra interventi di tutela epistemica e salvaguardia dei diritti fondamentali. La risposta a tali sfide richiederà un impegno multidisciplinare e forme di cooperazione istituzionale e internazionale più strette.

Nel complesso, il lavoro conferma che la sicurezza cognitiva non è un'estensione marginale della cybersecurity, ma una componente essenziale della resilienza democratica contemporanea. Rafforzarla significa dotare le società degli strumenti necessari per preservare la fiducia, la capacità decisionale e la coesione, in un ambiente informativo sempre più permeabile alla manipolazione e alla distorsione sistematica.

BIBLIOGRAFIA

AI4TRUST Consortium. (2023). *AI4Trust project overview (Horizon Europe Project 101070190)*. European Commission, Horizon Europe Programme. https://cordis.europa.eu/project/id/101070190

Amnesty International. (2022). *The Social Atrocity: Meta and the Right to Remedy for the Rohingya*. Amnesty International International Secretariat. Index: ASA 16/5933/2022.

Atlantic Council. (2020). *Iranian Digital Influence Efforts: Guerrilla Broadcasting for the Twenty-First Century* (E. T. Brooking & S. Kianpour). Digital Forensic Research Lab (DFRLab).

BBC News. (2019). *Fake voice used to scam company out of €220,000*. https://www.bbc.com/news/technology-48908736

Bhattacharya, S., & Singh, A. (2025). *Unravelling the infodemic: a systematic review of misinformation dynamics during the COVID-19 pandemic*. Frontiers in Communication, 10. https://doi.org/10.3389/fcomm.2025.1560936

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Oxford Internet Institute.

Bradshaw, S., & Howard, P. N. (2020). *Industrialized disinformation: 2020 global inventory of organised social media manipulation*. Oxford Internet Institute.

Council of the European Union. (2022). *Council Decision (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP...*. Official Journal of the European Union.

Cresci, S. (2020). *A decade of social bot detection*. Communications of the ACM, 63(10), 72–83. https://doi.org/10.1145/3409116

Cull, N. J., et al. (2017). *Soviet Subversion, Disinformation and Propaganda: How the West Fought Against it*. LSE Institute of Global Affairs.

Deshpande, V. (Ed.). (2018). *Hybrid Warfare: The Changing Character of Conflict*. Institute for Defence Studies & Analyses (IDSA) & Pentagon Press.

European Parliament and Council of the European Union. (2022). *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive)*. Official Journal of the European Union, L 333, 80–152.

Ferrara, E., et al. (2016). *The rise of social bots*. Communications of the ACM, 59(7), 96–104. https://doi.org/10.1145/2818717

Gallotti, R., et al. (2020). *Assessing the risks of "infodemics" in response to COVID-19 epidemics*. Nature Human Behaviour, 4(12), 1285–1293. https://doi.org/10.1038/s41562-020-00994-6

Goldstein, J. A., et al. (2023). *Generative AI for Information Operations: A Framework for Threat Assessment*. CSET, OpenAI, and Stanford Internet Observatory.

Goodfellow, I., et al. (2014). *Generative adversarial nets*. Advances in Neural Information Processing Systems, 27. https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Google Threat Analysis Group. (2023). *Over 50,000 instances of DRAGONBRIDGE activity disrupted in 2022*. Google Blog.

Hedling, E. (2025). *Social identities and democratic vulnerabilities: Learning from examples of targeted disinformation*. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE), Paper 24.

Helmus, T. C., & Chandra, B. (2024). *Generative Artificial Intelligence: Threats to Information Integrity and Potential Policy Responses*. RAND Corporation.

Hern, A. (2024, May 17). *UK engineering firm Arup hit by deepfake scam in Hong Kong*. The Guardian. https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video

Hoffman, F. G. (2007). *Conflict in the 21st Century: The Rise of Hybrid Wars*. Potomac Institute for Policy Studies.

Howard, P. N. (2020). *Lie machines: How to save democracy from troll armies, deceptive ads, and data ops*. Yale University Press.

Institute for Strategic Dialogue. (2025, March 18). *Investigation: How Russia Today is evading sanctions and spreading pro-Kremlin propaganda in Italy*. ISD Digital Dispatches.

International Organization for Standardization (ISO). (2022). *ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements*. ISO.

Iyengar, S., et al. (2019). *The origins and consequences of affective polarization in the United States*. Annual Review of Political Science, 22, 129–146.

Juurvee, I. (2018). *The resurrection of 'active measures': Intelligence services as a part of Russia's influencing toolbox*. Hybrid CoE Strategic Analysis 7.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kavanagh, J., & Rich, M. D. (2018). *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. RAND Corporation.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). *Beyond misinformation: Understanding and coping with the "post-truth" era*. Journal of Applied Research in Memory and Cognition, 6(4), 353–369.

Lewandowsky, S., et al. (2023). *Misinformation and the epistemic integrity of democracy*. Current Opinion in Psychology, 54.

Mandiant. (2022). *Dragonbridge: China-linked influence campaign*. Mandiant Blog.

Marushchak A., Petrov S., Khoperiya A. (2025), *Countering AI-powered disinformation through national regulation: learning from the case of Ukraine*, Frontiers in Artificial Intelligence, Vol. 7.

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society Research Institute.

NATO StratCom COE. (2021). *Social media manipulation 2021: State of the art*. NATO Strategic Communications Centre of Excellence.

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, U.S. Department of Commerce.

OECD (2024), *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity*, OECD Publishing, Paris.

Ó Fathaigh, R., et al. (2025). *The Regulation of Disinformation Under the Digital Services Act*. Media and Communication, 13.

Pennycook, G., & Rand, D. G. (2019). *Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning*. Cognition, 188, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux.

Roozenbeek, J., & van der Linden, S. (2019). *Fake news game confers psychological resistance against online misinformation*. Palgrave Communications, 5(65).

Sadek, D., & Mashkoor, L. (2023). *In Israel-Hamas conflict, social media become tools of propaganda and disinformation*. DFRLab Atlantic Council.

Sensity AI. (2020). *The State of Deepfakes: Landscape, Threats, and Impact*. Sensity AI.

Shao, C., et al. (2018). *The spread of low-credibility content by social bots*. Nature Communications, 9(4787).

Staves, A., et al. (2020). *An Analysis of Adversary-Centric Security Testing within Information and Operational Technology Environments*. Lancaster University.

Transparency International. (2024). *Fake news, corruption and compliance in the private sector*. Transparency International Helpdesk.

United States Department of State – Global Engagement Center. (2022). *RT and Sputnik's role in Russia's disinformation and propaganda ecosystem*. U.S. Department of State.

Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online*. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe Report.

World Health Organization. (2022). *Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating misinformation*. WHO EPI-WIN.

Zednik, C. (2019). *Solving the black box problem: A normative framework for explainable artificial intelligence*. Philosophy & Technology, 33, 491–518.