Malicious use of Artificial Intelligence (MUAI): l'uso malevolo dell'intelligenza artificiale nell'ecosistema cyber-sociale

di *Arije Antinori*[*]

La nascita ed evoluzione dell'Uso Malevolo dell'Intelligenza Artificiale (MUAI) rappresenta un cambio paradigmatico nell'ambito dei comportamenti cyberdevianti e cybercriminali, in termini di complessità, sofisticazione e velocità di diffusione tra le diverse categorie di attori dell'ecosistema cyber-sociale. L'adozione di strumenti e tecnologie di IA comporta il mutamento della natura stessa delle minacce alla sicurezza pubblica e nazionale, introducendo nuovi rischi e vulnerabilità. Il MUAI può essere utilizzato per generare messaggi, profili e utenti sintetici altamente realistici, contribuendo alla continua espansione dell'universo bot. Questo rende possibile, inoltre, la raccolta e analisi massiva di interazioni e dati personali sulle piattaforme sociomediali, permettendo la creazione di attacchi personalizzati e difficilmente rilevabili. Ciò consente di attivare operazioni di influenza, ingerenza e interferenza dello spazio pubblico, soprattutto politico, economico e sociale.

Parole chiave: IA; MUAI; sicurezza; deepfake; disinformazione; cybersecurity.

The malicious use of Artificial Intelligence (MUAI) in the cyber-social ecosystem

The emergence and evolution of Malicious Use of Artificial Intelligence (MUAI) represents a paradigm shift in cyber-deviant and cybercriminal behaviour in terms of complexity, sophistication and speed of diffusion among the various categories of actors in the cyber-social ecosystem. The adoption of AI tools and technologies is changing the very nature of threats to public and national security, introducing new risks and vulnerabilities. MUAI can be used to generate highly realistic synthetic messages, profiles and users, contributing to the continuous expansion of the bot universe. This also makes it possible to collect and analyse massive amounts of interactions and personal data on social media platforms, allowing for the creation of personalised and difficult-to-detect attacks. This enables operations of influence, interference and meddling in the public sphere, especially in the political, economic and social spheres.

Keywords: AI; MUAI; security; deepfake; disinformation; cybersecurity.

DOI: 10.5281/zenodo.17297557

[*] Sapienza Università di Roma. arije.antinori@uniroma1.it.

1. La natura disruptive dell'intelligenza artificiale

L''intelligenza artificiale (IA) è definita una tecnologia disruptive, ossia portatrice di un'innovazione dirompente in grado di rivoluzionare non solo il mercato, ma di imprimere un mutamento radicale nell'intera società. L'IA è in grado di automatizzare i processi complessi, riducendo la necessità di intervento umano in attività che fino ad oggi richiedevano tempo ed esperienza. Tale tecnologia cambia i modelli di business costringendo le aziende tradizionali a ripensare il loro modo di operare al fine di mantenere adeguati livelli di competitività. Ciò determina una riconversione delle competenze e la rapida sostituzione di alcune attività lavorative, nonché la creazione di professioni del tutto nuove connesse in particolare agli aspetti di gestione e sviluppo dell'IA. Alcuni strumenti tradizionali divengono obsoleti e mutano anche i sistemi di interazione con i *device*, come nel caso del riconoscimento vocale che affianca e in alcuni contesti sostituisce la digitazione. Ulteriore esempio di portata trasformativa di tale tecnologia è l'utilizzo di modelli che utilizzano il *Deep Learning* (DL) per la generazione di testi, come nel caso dei *Generative Pre-trained Transformer* (GPT), tra cui l'ormai diffuso interfaccia AI conversazionale *ChatGPT*, che stanno sostituendo, tra l'altro, le ricerche per mezzo dei motori di ricerca, il c*opywriting*, le traduzioni e i servizi di assistenza online, facilitando così gli utenti nella loro quotidianità non solo in ambito professionale. La capacità di imitare il comportamento intelligente umano, di fornire strumenti predittivi e analitici, l'automazione avanzata dei processi, la riduzione dei costi operativi e l'integrazione trasversale stanno determinando la trasformazione di interi settori, dall'intrattenimento alla sicurezza, dalla salute all'educazione. Tuttavia, per gli utilizzatori, tale innovazione può comportare sfide complesse, poiché occorre costantemente adattarsi a nuove funzionalità, gestire i rischi legati a eventuali vulnerabilità, ma soprattutto affrontare l'impatto che l'adozione accelerata dell'IA, soprattutto in contesti privi di un adeguato framework normativo, può determinare sull'intera società. I rischi principali derivano dalla continua evoluzione dei sistemi di IA, che possono introdurre nuove forme di dipendenza tecnologica, riducendo la capacità critica degli individui e delle organizzazioni nell'interpretare le informazioni e nel *decision making*, sempre più tecnologicamente assistito. Inoltre, la progressiva automatizzazione di processi chiave può generare problemi di equità nell'accesso alle opportunità di impiego, poiché la richiesta di competenze specifiche potrebbe superare la capacità di adattamento del mercato del lavoro. Inoltre, il processo di iterazione continua che caratterizza lo sviluppo e rilascio di tecnologia AI comporta vulnerabilità che includono la possibilità di esposizione a minacce *cyber*

sempre più sofisticate, dovute all'adattabilità e all'apprendimento delle macchine. La velocità con cui i sistemi AI vengono aggiornati può portare a criticità algoritmiche "interne", è il caso dell'*Artificial Inte[glitch]ence* (Antinori, 2019), e falle nella sicurezza difficili da rilevare tempestivamente, aumentando i rischi di attacchi *cyber* mirati e violazioni della *privacy*. Infine, la crescente dipendenza da sistemi algoritmici che producono decisioni autonome pone interrogativi cogenti su trasparenza, responsabilità e verificabilità delle informazioni generate.

2. L'Uso Malevolo dell'Intelligenza Artificiale

La diffusione dei sistemi AI ha permesso la riduzione dei costi di attacchi informatici e fisici, consentendo l'automazione di operazioni che in passato richiedevano l'intervento umano. Tale processo ha comportato un'importante espansione del bacino di attori malevoli in grado di condurre attacchi, una maggiore frequenza delle azioni offensive e una dilatazione del piano d'attacco, ossia una crescita esponenziale del numero di potenziali *target*. Per Uso Malevolo dell'Intelligenza Artificiale (MUAI) si intende generalmente l'utilizzo di tecnologie AI per il conseguimento di diversi obiettivi dannosi, non etici e/o illeciti (Pashentsev, 2023). Appare evidente, quindi, come esso si sia sviluppato parallelamente all'evoluzione stessa dell'IA (Brundage *et al.,* 2018). Il MUAI include, dal punto di vista criminale, un ampio spettro di attività che vanno dalle intrusioni nei sistemi informatici alla realizzazione di contenuti e campagne disinformative, dall'hackeraggio di sistemi militari autonomi alla compromissione dell'integrità di tecnologia AI, da modalità avanzate di *social engineering* (Blauth *et al.,* 2022) alla generazione di contenuti sintetici. Il MUAI si manifesta prevalentemente in ambiti quali il cybercrime, la cyberdeviance, intesa come l'insieme di pratiche cyber-sociali che violano e/o compromettono norme sociali e/o giuridiche, o ne contestano i confini, generando al contempo processi di adesione, seppur limitati asingoli e gruppi, e reazione estesa da parte dell'opinione pubblica, in un contesto in cui la devianza – tradizionalmente intesa – ne risulta amplificata quindi in termini di scala, velocità e tracciabilità, coinvolgendo attori umani e non-umani. Il MUAI risulta, inoltre, particolarmente efficace nell'ambito della cyberwar e della manipolazione dell'opinione pubblica, trasformando in modo significativo tanto il panorama della *cybersecurity* quanto quello della sicurezza cyber-sociale (Antinori, 2018). A livello strategico, si individuano due macroaree di applicazione, una logica e una sociale.

La prima riguarda l'evoluzione e il potenziamento dei cyberattacchi tradizionali, sfruttando le vulnerabilità logiche dei sistemi informatici e delle componenti *hardware*, al fine di compromettere confidenzialità, integrità e disponibilità dei dati. I cybercriminali ricorrono al MUAI per automatizzare e rendere più efficaci i loro attacchi, come nel caso del *phishing* avanzato o del *code injection*, principalmente a danno di apparecchi e veicoli a guida autonoma, nonché gli attacchi ai sistemi di autenticazione biometrica. Il MUAI consente altresì di sviluppare *malware* adattivi, in grado di evolversi e aggirare le tradizionali misure di sicurezza, e lo sviluppo di concetti teorici per ransomware basati su Reinforcement Learning (RL), come nel proof-of-concept accademico denominato RansomAI, che dimostra come un agente potrebbe apprendere *ransomware* basati su RL, come *RansomAI* che utilizza un agente per apprendere in tempo reale quale algoritmo di crittografia, velocità e durata adottare al fine di minimizzare la possibilità di essere individuato, massimizzando al contempo il danno inflitto (von der Assen *et al.,* 2023). Inoltre, le tecniche di *Machine Learning* (ML) consentono di individuare e sfruttare vulnerabilità nei *software* con una rapidità mai vista prima, aumentando così il rischio di cosiddetti *zero-day attack*, spesso volti alla compromissione su ampia scala delle infrastrutture critiche. In tale macroarea, pertanto, il MUAI ha determinato:

- potenziamento delle minacce già esistenti;
- ridefinizione sostanziale delle caratteristiche delle minacce;
- sviluppo di nuove tipologie di attacchi;
- introduzione di nuove minacce attraverso strategie/tattiche in grado di sfruttare anche le vulnerabilità dei sistemi difensivi basati sull'IA.

Ciò rende possibili attacchi prima impensabili in quanto a complessità e risorse necessarie, favorendo la proliferazione di strategie, metodologie, tecniche e tattiche offensive, caratterizzate da un elevato livello di precisione, in grado di superare le tradizionali difese, rendendo sempre più difficili il rilevamento e l'attribuzione. Così, l'automazione dei cyberattacchi potenziati, in termini di scalabilità e velocizzazione, grazie all'uso di strumenti *AI-driven*, pone importanti sfide alla *cybersecurity*. Gli attori malevoli, sia simmetrici che asimmetrici, tra cui: potenze ostili, organizzazioni criminali, mafie, entità estremistico-violente, terroristi, cybercriminali e *hacking crew*, possono oggi grazie al MUAI, compromettere la sicurezza, e non solo, di individui, aziende, comunità e governi (UNOCT, 2021), contribuendo a modificare il tradizionale gap asimmetrico operativo tra attaccante e attaccato. Sebbene alcuni attori possano ridurre tale gap, altresì sfruttando la capacità trasformativa *dual-use* di tecnologie AI, altri esperti evidenziano il rischio

che l'IA possa al contrario ampliarlo a favore degli attori con maggiori risorse e capacità difensive. Pertanto, occorre sviluppare linee guida di natura algoretica al fine di prevenire degenerazioni criminali del dual-use nell'IA (Brundage *et al.,* 2018). Tuttavia, occorre precisare che il MUAI non deve essere confuso con l'*ethical hacking* che può servirsi di strumenti IA, nell'individuazione e disvelamento di sistemi AI corrotti o compromessi, a condizione che le attività connesse siano condotte entro certi limiti dal punto di vista legale (Choraś, Woźniak, 2021).

La seconda macroarea di applicazione del MUAI riguarda il potenziale sfruttamento delle "vulnerabilità sociali", come esito dell'interazione fra caratteristiche individuali, quali ad esempio l'età (Smahel *et al.,* 2020), appartenenze di gruppo (Abrajano *et al.,* 2024), e specificità algoritmiche delle diverse piattaforme sociomediali, siano esse *mainstream* o *fringe*, in quanto non più meri contenitori neutri, ma agglomerati connettivi di pubblici umani e inumani (Győri *et al.,* 2022). Il MUAI può quindi amplificare dinamiche di omofilia e polarizzazione affettiva all'interno delle cosiddette *echo chambers* – spazi informativi in cui le voci dissonanti sono ridotte al minimo, se non del tutto assenti, screditate e/o marginalizzate –, nonché l'esposizione selettiva rafforza credenze preesistenti, sfruttando ripetizione, novità ed emozioni per innescare elaborazione euristica e ragionamento motivato, all'interno di infosfere in cui le *affordance* di piattaforma contribuiscono a moltiplicare la portata dei contenuti manipolativi. Pertanto, vulnerabilità individuali e strutture algoritmiche si co-producono, trasformando così rischi informativi in minacce alla sicurezza cyber-sociale, quindi inevitabilmente politica, economica, e sempre più pubblica, nazionale e globale. In sintesi, le "vulnerabilità sociali", oggetto di interesse del MUAI, non si riducono a tratti individuali, ma emergono dove bias cognitivi, identità di gruppo e affordance di piattaforma si sovrappongono a fragilità, scarsa capacità o assenza di governance e moderazione. In tal senso, il MUAI riulta particolarmente efficace per la manipolazione psicologica e comportamentale degli individui attraverso l'automazione della disinformazione, creazione di *deepfake*, uso di *cyber-bot* e *AI-bot* per il *trolling* volto a influenzare il dibattito pubblico e favorire la disseminazione di narrazioni tossiche e divisive, la polarizzazione e weaponizzazione delle *audience* (Győri *et al.,* 2022). Tale categoria di minacce è particolarmente pericolosa in quanto sfrutta meccanismi cognitivi e dinamiche cyber-sociali al fine di alterare la percezione della realtà, indebolire la fiducia nelle istituzioni e creare divisioni tra gruppi sociali. Le campagne di disinformazione basate sul MUAI possono diffondere narrazioni manipolate con una velocità, una targetizzazione e una dimensione di scala impensabili fino a pochi anni fa, rendendo estremamente complessa la loro

individuazione, mitigazione e neutralizzazione. Le azioni offensive basate sul MUAI si distinguono per un'elevata precisione e una maggiore resilienza ai tentativi di contrasto, il che rende piuttosto complicata sia la difesa che la prevenzione, così come il contrasto e la mitigazione delle seguenti finalità operative delle tattiche MUAI:

- *deception* – creazione per mezzo della *Generative AI (GenAI)* di contenuti falsi, come testi, immagini, audio/video, volti a ingannare individui e istituzioni (Park *et al.,* 2024);
- manipolazione sociale – utilizzo di *bot* e algoritmi per influenzare l'opinione pubblica e diffondere propaganda (Marcellino *et al.,* 2023);
- furto di dati e violazioni della privacy – intrusioni abusive nei sistemi e raccolta massiva di informazioni e/o dati sensibili per campagne di *doxing* (Khalid *et al.,* 2023);
- frode finanziaria – frodi, in particolare su carte di credito e cryptovalute, evasione e/o elusione automatizzata dei sistemi di rilevamento antifrode, nonché manipolazione dei mercati (Josyula *et al.,* 2023);
- *deepfake* e disinformazione – creazione e diffusione di contenuti sintetici realistici con l'obiettivo di influenzare decisioni politiche ed economiche, nonché manipolare l'opinione pubblica con impatti significativi a livello psicologico, sociale e politico (Kharvi, 2024).

Se il crescente impiego di sistemi AI nella quotidianità delle società più avanzate da un lato favorisce la fruizione di servizi e l'automazione di procedure in termini di efficienza a favore del cittadino, dall'altro facilita la progettazione e realizzazione di operazioni malevoli contro *target* specifici, riducendo il margine di errore e amplificando l'impatto delle azioni criminali.

3. La contaminazione sintetica dell'ecosistema cyber-sociale

Internet, ciò che abbiamo definito in principio, spazio virtuale, evidenziandone la dimensione pseudo-ludica, immateriale, ma soprattutto non-reale, e poi cyberspace, ossia rete globale di computer interconnessi, secondo una prospettiva interpretativa fortemente tecnocentrica, rappresenta oggi, grazie all'evoluzione infrastrutturale digitale – nella transizione dai new media ai social media, attraverso la diffusione online di piattaforme algoritmiche ininterrottamente abitate da milioni di utenti nel mondo –, e al rapido mutamento tecnosociale, un ecosistema cyber-sociale. Questo deve essere

inteso oggi come un sistema socio-tecnico multidimensionale caratterizzato non solo da infrastrutture e piattaforme, ma da attori umani e non-umani, istituzioni, pratiche culturali e mediali, fenomeni sociali, e framework normativi, che influenza in modo significativo tanto la sicurezza quanto la coesione sociale.

In tale contesto, l'implementazione accelerata dell'AI, in assenza di un adeguato controllo istituzionale, può favorire la polarizzazione del dibattito pubblico, l'accentuazione delle disuguaglianze economiche, la proliferazione e disseminazione di contenuti ingannevoli. Così la creazione di *deepfake* e identità sintetiche, quindi false, rappresenta una delle principali risorse del MUAI in termini di costi ridotti e flessibilità operativa, in particolare attraverso la disseminazione massiva nell'ecosistema cyber-sociale di disinformazione, propaganda estremistica, odio online, al fine di attaccare e/o compromettere la reputazione di individui e organizzazioni, rappresentare sinteticamente *leader* politici raffigurati mentre dicono o compiono azioni di fatto mai accadute, influenzando elezioni e destabilizzando governi (Pashentsev, 2021).

Il *deepfake* costituisce una delle evoluzioni più sofisticate dell'AI applicata all'elaborazione di immagini e video, mentre nel caso della sintesi o clonazione della voce umana si hanno i *deepaudio*. Tecnologie come le *Generative Adversarial Networks* (GANs) consentono di creare contenuti audio/video altamente realistici attraverso l'uso di reti neurali e modelli generativi, permettendo tra l'altro la sostituzione di volti originali con quelli di altre persone, al fine di ottenere un effetto estremamente convincente. Il processo di creazione del *deepfake* si basa sull'addestramento AI per mezzo di un numero molto elevato di immagini del volto *target*, spesso reperite attraverso le piattaforme sociomediali. Occorre precisare che tale prodotto mediale differisce dal *deep video portrait*, che non comporta una sostituzione del volto, ma solo una manipolazione delle espressioni di un attore dal vero. Inoltre, il *deepfake* si inserisce in un quadro più ampio di manipolazione che comprende gli *shallow fakes*, ossia video alterati in modo grossolano e superficiale – ad esempio, attraverso un cambio della velocità e/o una manipolazione dell'audio – e i cosiddetti *synthetic media*, contenuti interamente generati dall'IA, come gli *avatar* iperrealistici. I primi *deepfake* sono apparsi sulla piattaforma *Reddit*, nella forma di contenuti di natura pornografica che vedevano la sovrapposizione dei volti di celebrità a quelli di pornoattori, al fine di realizzare video artefatti di natura sessuale. In tale contesto, la facilità di accesso ad app mobile e la disponibilità online di tool AI ha favorito una rapida diffusione del fenomeno, con migliaia di utenti che hanno iniziato a produrre e condividere contenuti. Negli ultimi anni, è aumentata

significativamente la capacità di manipolare contenuti audiovisivi, soprattutto negli ambiti politici e militari, portando alla maturazione di un mercato clandestino di *deepfake_as_a_service*, che consente anche a individui privi di competenze tecniche di accedere a strumenti avanzati per la creazione di contenuti falsificati (Europol, 2022). Tale fenomeno non solleva preoccupazioni soltanto per quanto concerne la prospettiva tradizionale di *cybersecurity*, ma sempre più in relazione alla disinformazione e manipolazione dell'opinione pubblica nell'ecosistema cyber-sociale in un contesto di post-verità (Antinori, 2019). L'uso indiscriminato di tale tecnologia per la personalizzazione dei contenuti informativi può favorire e accelerare la creazione di *echo chambers*, in particolare rafforzando pregiudizi sulla base di bias cognitivi. I contenuti *deepfake* possono essere impiegati come strumenti di propaganda, sabotaggio e *cyberwarfare*, alimentando teorie del complotto, quindi favorendo l'azione di gruppi cospirazionistici, estremismi antisistema e antigoverno (ASAGe), nonché fomentando l'odio *online*. Nel 2024, considerato come "*global elections year*" per le importanti elezioni che si sono succedute in diversi continenti, tra cui quelle statunitensi, il ricorso ai *deepfake* per alterare le dichiarazioni di candidati politici ha minato la fiducia pubblica e diffuso disinformazione su ampia scala (Insikt Group, 2024). Il MUAI trova la sua applicazione più pervasiva e dannosa a livello individuale nella creazione di pornografia non consensuale, che costituisce la stragrande maggioranza dei contenuti deepfake online. Questo si aggiunge ad altri usi criminali come il *revenge porn*, la pornografia non consensuale e la pedopornografia *online*, con l'obiettivo di danneggiare gravemente la reputazione degli individui, di celebrità, di importanti personaggi politici, ma soprattutto colpendo gli sceenagers, adolescenti e pre-adolescenti sempre connessi (Antinori, 2021) Risulta evidente come la capacità MUAI di falsificare prove video e audio ponga nuove sfide in ambito giudiziario, richiedendo l'adozione di metodologie e tecniche avanzate per garantire la custodia e genuinità delle prove digitali acquisite nell'ambito dei procedimenti penali (Di Paolo, Pressacco, 2022). La tecnologia *deepfake* sta rapidamente evolvendo, rendendo sempre più difficile distinguere tra contenuti autentici e manipolati, costituendo altresì una minaccia alla sicurezza nazionale. Tale aspetto risulta particolarmente rilevante in scenari di crisi e/o conflitti geopolitici, tanto da divenire una modalità strutturata nell'ambito della *Cognitive Warfare* (Antinori, 2024).

Conclusioni

Lo sviluppo e diffusione del MUAI segna un punto di svolta nell'ambito della sicurezza globale, imponendo nuove sfide alla capacità di difesa e resilienza delle società contemporanee. La crescente pervasività di tale tecnologia dimostra come la minaccia non sia più confinata al solo ambito della *cybersecurity*, ma si estenda in modo trasversale all'intero ecosistema cybersociale, influenzando dinamiche politiche, economiche e culturali. In particolare, la capacità del MUAI di manipolare l'opinione pubblica, sfruttare vulnerabilità psicologiche, sociali, e alterare la percezione della realtà attraverso *deepfake* e campagne disinformative evidenzia l'urgenza di strategie di contrasto integrate, multidisciplinari e multiattoriali.

In tale contesto, la regolamentazione e la *governance* dell'AI assumono un ruolo cruciale. Risulta necessario sviluppare un framework normativo convergente a livello globale che disciplini l'AI, al fine di evitare il proliferare di strumenti sempre più sofisticati nelle mani delle diverse tipologie di attori malevoli. Da qui la necessità di trovare un equilibrio tra innovazione e protezione della sicurezza, sia pubblica che nazionale, favorendo il progresso tecnologico, ma mitigandone i rischi attraverso misure di controllo efficaci e trasparenti. Il rafforzamento della cooperazione internazionale tra governi, università, enti di ricerca e aziende private risulta imprescindibile per prevenire l'abuso delle tecnologie AI e promuovere standard condivisi per l'identificazione e neutralizzazione di contenuti manipolati.

Risulta altresì essenziale un'azione mirata alla formazione, sensibilizzazione e consapevolezza a livello sociale. L'educazione digitale e la *AI-literacy*, devono includere la capacità di riconoscere la disinformazione e comprendere i meccanismi di manipolazione utilizzati nel contesto cybersociale con l'obiettivo di rafforzare la fiducia pubblica nelle istituzioni, la coesione sociale e la stabilità democratica.

Riferimenti bibliografici

Abrajano M., Garcia M., Pope A., Vidigal R., Tucker J.A., Nagler J. (2024). How reliance on Spanish-language social media predicts beliefs in false political narratives amongst Latinos. *PNAS Nexus*, 3(11): pgae442. https://doi.org/10.1093/pnasnexus/pgae442

Anderson R., Barton C., Böhme R., Clayton R., Van Eeten M., Levi M. (2019). Measuring the cost of cybercrime. In: Moore T., Pym D., Ioannidis C. (a cura di), *The Economics of Information Security and Privacy* (pp. 265-300). Cham: Springer.

Antinori A. (2018). *Artificial Intelligence. Una minaccia sistemica alla sicurezza nazionale in prospettiva futura*, *GNOSIS – Rivista Italiana di Intelligence*, AISI – Agenzia Informazioni e Sicurezza Interna.

Antinori A. (2018). Sicurezza cyber-sociale. In: Baldoni R., De Nicola R., Prinetto P. (a cura di), *Il futuro della cybersecurity in Italia: Ambiti progettuali strategici* (pp. 188-190). Roma: Laboratorio Nazionale di Cybersecurity, CINI.

Antinori A. (2019). Terrorism and deepfake: From hybrid warfare to post-truth warfare in a hybrid world. In: Griffiths P., Kabir M.N. (a cura di), *Proceedings of the European Conference on the Impact of AI and Robotics* (pp. 23-30). Reading: Academic Conferences and Publishing International Limited.

Antinori A. (2020). Terrorism in the early onlife age: From propaganda to 'propulsion'. In: Antinori A. (a cura di), *Terrorism and Advanced Technologies in Psychological Warfare. New Risks, New Opportunities to Counter the Terrorist Threat*. New York: Nova Science Publishers.

Antinori A. (2021). La sicurezza cyber-sociale dei giovani (e non solo). Tra social network, metaversi e radicalizzazione. *CyberSecurity Italia*. https://www.cybersecitalia.it/la-sicurezza-cyber-sociale-dei-giovani-e-non-solo-tra-social-network-metaversi-e-radicalizzazione/15171/

Antinori A. (2024). La cognitive warfare in uno scenario di minacce convergenti. *Sicurezza, Terrorismo e Società*, 19(Special Issue 1/2024): 67-81. Milano: EDUCatt – Università Cattolica del Sacro Cuore.

Bazarkina D., Pashentsev Y.N. (2019). Artificial intelligence and new threats to international psychological security. *Russia in Global Affairs*.

Biggio B., Nelson B., Laskov P. (2012). Poisoning attacks against support vector machines. arXiv:1206.6389.

Blauth T.F., Gstrein O.J., Zwitter A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, 10: 77110-77122. https://doi.org/10.1109/ACCESS.2022.3191790

Brown A.W. (2021). *Surviving and Thriving in the Age of AI. A Handbook for Digital Leaders*. Cambridge: Cambridge University Press.

Brundage M., Avin S., Clark J. et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv*. https://doi.org/10.48550/arXiv.1802.07228

Carlini N., Wagner D. (2017). Towards evaluating the robustness of neural networks. arXiv:1608.04644.

Choraś M., Woźniak M. (2021). The double-edged sword of AI: Ethical adversarial attacks to counter artificial intelligence for crime. *SN Computer Science*, 2: 1-4. https://doi.org/10.1007/S43681-021-00113-9

Di Paolo G., Pressacco L. (a cura di) (2022). *Intelligenza artificiale e processo penale. Indagini, prove, giudizio*. Trento: Università degli Studi di Trento.

Europol (2022). *Facing reality? Law enforcement and the challenge of deepfakes*. The Hague: Europol Innovation Lab. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcementand-challenge-of-deepfakes

Ferrara E. (2023). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. arXiv:2310.00737.

Floridi L., Cowls J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

Frey C.B. (2019). *The Technology Trap. Capital, Labor, and Power in the Age of Automation*. Princeton: Princeton University Press.

Goodfellow I.J., Shlens J., Szegedy C. (2015). Explaining and harnessing adversarial examples. arXiv:1412.6572.

Gu T., Dolan-Gavitt B., Garg S. (2019). BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv:1708.06733.

Győri L., Molnár C., Krekó P., Szicherle P. (2022). The Kremlin's troll network never sleeps: Inauthentic pro-Kremlin online behavior on Facebook in Germany, Italy, Romania and Hungary. Budapest: Political Capital. https://politicalcapital.hu/pc-admin/source/documents/pc_ned_study_kremlin_troll_network_2022_web.pdf

Insikt Group (2024, 24 settembre). Targets, objectives, and emerging tactics of political deepfakes. *Recorded Future*. https://www.recordedfuture.com/research/targets-objectives-emerging-tactics-political-deepfakes

Jansevskis M., Osis K. (2023). *Artificial Intelligence for Security. Enhancing Protection in a Digital Age*. Cham: Springer Nature.

Josyula H.P., Vishnubhotla D., Onyando P.O. (2023). Is artificial intelligence an efficient technology for financial fraud risk management?. *International Journal of Managerial Studies and Research (IJMSR)*, 11(6): 11-16. https://doi.org/10.20431/2349-0349.1106002

Khalid N., Qayyum A., Bilal M., Al-Fuqaha A., Qadir J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158: 106848. https://doi.org/10.1016/j.compbiomed.2023.106848

Kharvi P. (2024). Understanding the impact of AI-generated deepfakes on public opinion, political discourse, and personal security in social media. *IEEE Security & Privacy*, 22(2): 2-9. https://doi.org/10.1109/MSEC.2023.106848

Kurakin A., Goodfellow I., Bengio S. (2017). Adversarial examples in the physical world. arXiv:1607.02533.

Marcellino W., Beauchamp-Mustafaga N., Kerrigan A., Chao L.N., Smith J. (2023). The rise of generative AI and the coming era of social media manipulation 3.0: Next-generation Chinese astroturfing and coping with ubiquitous AI. Santa Monica (CA): RAND Corporation. https://www.rand.org/pubs/perspectives/PEA2679-1.html

Marchal N., Xu R., Elasmar R., Gabriel I., Goldberg B. (2024). Generative AI misuse: A taxonomy of tactics and insights from real-world data. arXiv:2406.13843.

Masayuki I. (2021). *A Narrative History of Artificial Intelligence. The Perpetual Frontier of Information Technology*. Cham: Springer Nature.

Papernot N., McDaniel P., Wu X., Jha S., Swami A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv*:1511.04508.

Park P.S., Goldstein S., O'Gara A., Chen M., Hendrycks D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5): 100988. https://doi.org/10.1016/j.patter.2024.100988

Pashentsev E. (2021). The Palgrave handbook of malicious use of AI and psychological security (pp. 45-60). London: Palgrave Macmillan.

Pashentsev E. (a cura di) (2023). *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-22552-9

Political Capital (2022). *The Kremlin's Troll Network Never Sleeps. Inauthentic Pro-Kremlin Online Behavior on Facebook in Germany, Italy, Romania, and Hungary*. Budapest: Political Capital.

Smahel D., Machackova H., Mascheroni G., Dedkova L., Staksrud E., Ólafsson K., Livingstone S., Hasebrink U. (2020). *EU Kids Online 2020. Survey results from 19 countries*.