UFV - gruppo di lavoro – manipolazione dello spazio pubblico
Note preparatorie - 15 dicembre – NATO CCDCOE

Questioni da discutere

- Considerando lo sviluppo delle recenti tecnologie e la tua esperienza sul campo, come è possibile risalire alle impronte FN fino alla loro origine?
- le impronte FN costituiscono una traccia permanente (come nel caso delle Block-Chain) o sono immediatamente o rapidamente cancellabili e alterabili?
- I protocolli di verifica sono idonei a garantire l'integrità delle informazioni a fini probatori?
- Su quali basi si può differenziare una campagna di disinformazione dall'orientamento (legittimo) e dall'influenza dei media sull'opinione pubblica?
- Sono stati sviluppati parametri per classificare una diffusione di notizie come parte di una campagna di disinformazione?
- In che modo l'IA facilita le campagne di disinformazione in un ambiente virtuale? Vi sono stati casi concreti di utilizzo di strumenti di intelligenza artificiale per questo scopo?
- Le tecnologie di IA sono state utilizzate come strumento efficace per prevenire e contrastare la disinformazione?
- Come è possibile verificare l'efficacia delle campagne di disinformazione nell'influenzare l'opinione pubblica?
- Sono stati segnalati casi in cui questo specifico tipo di campagna è stato oggetto di indagini penali? In caso affermativo, quali questioni sono state affrontate? Quali problemi sono sorti nella cooperazione internazionale?
- In generale, quanto sono affidabili gli strumenti attualmente disponibili per l'attribuzione? Qual è la rilevanza dei dati tecnici/digitali nel processo di attribuzione ai sensi del diritto internazionale (ad esempio, la catena di passaggi ricostruiti nello spazio virtuale) rispetto a quelli non digitali (*cui prodest* - moventi; caratteri linguistici o grafici; orientamento politico del soggetto operante, ecc.)?
- Quali problemi giuridici sono sorti nella pratica e quali sono immaginabili?

Note esplicative

In questi due anni la FVO ha discusso di come l'IA stia cambiando il modo di guardare ai crimini transnazionali. Abbiamo pubblicato in un importante Rivista Penale on-line il workshop tenutosi nel novembre 2021; l'Interagency Law Enforcement Academy of Advanced Studies (un centro di formazione per tutti i corpi di polizia italiani, istituito presso il Ministero dell'Interno) sta per pubblicare nella propria Rassegna il contributo della FVO al Corso 2022, sugli aspetti di rilevanza penale della AI. Nel prossimo anno si terranno due Corsi Occorsio sulle stesse tematiche presso la Scuola Superiore per la Magistratura.

A nostro avviso, il risultato più importante raggiunto potrebbe essere considerato la consapevolezza circa la questione dell'applicazione della giurisdizione nel cyberspazio e le sue relazioni con problemi analoghi sorti nel campo del diritto internazionale. La questione della "**attribuzione**" può essere considerata molto vicina a quella della raccolta di prove nella sovranità di altri Stati. Di conseguenza, abbiamo discusso dell'efficacia del secondo protocollo aggiuntivo alla Convenzione di Budapest nel campo della **cooperazione giudiziaria** e dei risultati del manuale Tallinn II, dall'altro lato.

Ora la nostra attenzione è su come sia possibile (secondo i confini costituzionali e convenzionali) fornire **strumenti penali contro la disinformazione** nello spazio pubblico.

Abbiamo discusso l'ipotesi di contrastare l'uso delle Fake News, anche considerando l'azione penale. Le questioni giuridiche in gioco sono sufficientemente chiare. Abbiamo individuato una possibile linea di discussone, consistente nell'applicare vecchi strumenti, previsti dalle leggi esistenti, e nell'introdurre anche nuovi reati; tali ultime ipotesi volte a garantire la correttezza procedurale della distribuzione delle notizie (identificazione degli autori e della provenienza; identificazione dei BOT e così via).

Abbiamo raggiunto, allo stesso tempo, la consapevolezza della più ampia estensione della disinformazione. Le FN sono parte di un problema più complesso, e potrebbero non essere i più importanti o pericolosi.

Quindi, per completare il nostro lavoro sulla manipolazione dello Spazio Pubblico, come pericolo imminente per la fiducia come fondamento della democrazia, vorremmo raccogliere maggiori informazioni sul tema del Cognitive Warfare, nel contesto della Disinformazione. Quali sono le esperienze più recenti in questo campo? Come funziona l'uso dell'IA nel contesto della disinformazione? Quanto sono importanti le diverse forme di influenza? Come è possibile individuare una campagna di disinformazione e attribuirla a uno Stato o a una parte non statale? Il processo di attribuzione interferisce con la sovranità dello Stato e quali sono le questioni giuridiche in gioco? In quale direzione procede la revisione del Manuale di Tallinn su questi punti?

I seminari.

Il focus dei Seminari è sul ruolo della punizione penale nel contrastare la disinformazione come strumento di condizionamento dello spazio pubblico; In questo contesto, stiamo esplorando la necessità della previsione di nuove fattispecie di reato, volte a preservare i valori fondanti della democrazia, primo fra tutti il valore della fiducia.

Una risposta globale ha un maggiore effetto deterrente e potrebbe limitare il rischio di escalation. In questo contesto, potrebbe essere molto importante rendere efficace l'intervento criminale, come parte della deterrenza basata sul diritto internazionale. [1]

La tensione che ha portato al conflitto ucraino ha rifatto riemergere il tema della disinformazione, volta a condizionare lo spazio politico attraverso tecniche complesse, specificamente adattate agli obiettivi perseguiti e allo sviluppo delle tecnologie.

L'influenza sulle elezioni presidenziali negli Stati Uniti, 2016, si è basata essenzialmente su metodi intrusivi tradizionali (penetrazione in un sistema informatico e conseguente diffusione di documenti così esfiltrati). La diffusione di informazioni rubate è stata facilitata da vari strumenti, dai media alle piattaforme social.

Abbiamo poi assistito all'utilizzo di video e immagini realizzate tramite GAN (Generative Adversial Network) e NLP (Natural Language Processing), riproducendo anche personaggi pubblici reali. Questi strumenti sono stati utilizzati anche di recente nel conflitto ucraino. La credibilità di questo strumento è data dalla capacità del creatore-utente di coniugare la crescente verosimiglianza del prodotto, grazie allo sviluppo della tecnica GAN e NLP, e l'adeguatezza delle false informazioni allo specifico target.

Dai casi noti, però, ne consegue che l'utilizzo di questo genere di strumenti, che possiamo definire in breve come Fake News, è solo una parte e non necessariamente la più importante delle tecniche di manipolazione, che si adattano ai diversi contesti.

Il FN implica la costruzione di un fatto falso (ad esempio, una persona politica esistente che fa o afferma qualcosa; o la creazione di un soggetto inesistente ma plausibile, che diventa un interlocutore in un dato

[1] "L'aumento della portata e dell'intensità delle minacce ibride, combinato con il destabilizzante effetti delle nuove tecnologie, possono portare a un'escalation non intenzionale e imprevedibile" (CoE)

dominio pubblico). Pertanto, la logica del contrasto si basa sullo svelamento della falsità. Questo può accadere sia attraverso la de-costruzione tecnica del falso, sia attraverso una contro-narrazione. In entrambi i casi c'è un *fatto* isolato e la reazione può limitarsi a disconoscere la sua falsità, intesa come non corrispondenza. Dal punto di vista penale, tale situazione non è diversa dall'accertamento giudiziario di una notizia "falsa" nell'abuso di mercato, o in altri casi in cui la falsità di un fatto asserito è elemento costitutivo del reato.

Nel corso del nostro lavoro, tuttavia, sono sorti due problemi.

In primo luogo, la crescente potenzialità ingannevole di questi strumenti, dovuta al progresso tecnico.

In secondo luogo, la difficoltà di operare una contro-narrazione efficace, derivante dal progressivo utilizzo dell'AI non solo nella creazione del FN, ma nei meccanismi della sua diffusione: il ML e la capacità del programma di reagire autonomamente e immediatamente alla contro-narrazione.

Questi comportamenti dovrebbero essere considerati come parte di tecniche più ampie di disinformazione o manipolazione dello spazio pubblico.

Sulla base delle fonti internazionali, consideriamo la "disinformazione" un'azione strategica deliberata, con l'obiettivo di disorientare e dividere l'opinione pubblica, polarizzarla, delegittimare le istituzioni e mettere in discussione l'idea stessa di democrazia[2]. Siamo consapevoli dell'importanza di coprire l'intero spettro della vita sociale. Pertanto, siamo interessati a radicare nel nostro specifico campo di interesse, ciò che il CoE - Centre of Excellence for Hybrid Threats (marzo 2022) ha definito come le implicazioni rivoluzionarie dell'IA: quando si muove oltre la manipolazione delle decisioni umane nella logica imperscrutabile delle macchine intelligenti.[3]

È ormai evidente che le tecniche di disinformazione si basano su più livelli e con strumenti diversi, spesso utilizzati da anni e che combinano aspetti cyber (come l'acquisizione di informazioni reali attraverso la penetrazione o l'effettiva costruzione di FN, nei termini sopra indicati; o ancora la rottura dei sistemi informatici, allo scopo di screditare l'istituzione responsabile) ad altri che agiscono nel mondo non virtuale e che sfruttano narrazioni politiche e strumenti di comunicazione di origine reale.

Esempi di questo adattamento possono essere gli eventi italiani e francesi. In entrambi, sembra esserci un legame tra la diffusione delle narrazioni NO-VAX e la successiva diffusione di narrazioni filo-russe contro il sostegno internazionale all'Ucraina.

L'adeguatezza delle tecniche di manipolazione al bersaglio deve considerare le caratteristiche specifiche della popolazione bersaglio e la resilienza del sistema politico attaccato, inteso nel suo complesso.[4]

Il punto è, per gli obiettivi della nostra ricerca, di notevole importanza.

Mentre è possibile individuare una FN e, di conseguenza, immaginare l'uso dello strumento penale per contrastarla (anche a fini deterrenti), nel rispetto dei principi costituzionali e convenzionali, sembra molto più difficile individuare nelle narrazioni e nella loro diffusione un "fatto" che possa costituire la base oggettiva di un reato.

[2] Come parte di Cognitive Warfare, cercando di "seminare dubbi, introdurre narrazioni contrastanti, polarizzare opinioni, radicalizzare e motivarli ad atti che possono distruggere una società altrimenti coesa" nelle parole di un pool di studiosi della John Hopkins Un. e Imperial College.

[3] Citando Payne

[4] Va considerato che la resilienza, nella deterrenza della disinformazione, è anche una sfida che può minare le fondamenta della società liberale, colpendo "le pietre angolari della democrazia occidentale: intervento statale limitato, pluralismo, media liberi e apertura economica".

Entrambe le ipotesi (FN e narrativa) sollevano anche la questione dell'attribuzione della condotta a soggetti specifici (persone, organizzazioni o stati; organizzazioni statali o non statali).

L'attribuzione pone due problemi distinti.

Il primo è il grado di certezza. In questa direzione, l'acquisizione di prove tecniche che portano all'attore è di grande – decisiva – importanza. Queste prove devono soddisfare standard diversi, a seconda che siano destinate alla prova processuale o alla responsabilità politica in conformità con il diritto pubblico internazionale (IL).

In entrambi i casi, queste prove devono anche soddisfare standard procedurali che ne garantiscano la conservazione a fini probatori, naturalmente con i diversi standard applicabili a fini probatori o per la responsabilità politica dello Stato.

L'acquisizione di queste certezze tecniche avviene necessariamente, almeno in parte, nel quadro della sovranità di altri Stati. Questi Stati possono essere neutrali rispetto a quello da cui ha effettivamente origine l'attacco.

Questo è il secondo problema. La violazione della sovranità di altri Stati si pone in termini diversi nel diritto penale interno e in IL.

Nel primo caso, può portare all'illegittimità in giudizio delle prove ottenute. In alcuni casi, può portare all'affermazione dell'illegalità dell'azione delle autorità pubbliche. In Italia la questione è già stata sollevata, dal momento che coloro (tra cui il pubblico ministero e gli ufficiali di polizia giudiziaria) che avevano "seguito" la traccia dell'attacco informatico al server, situato in un paese straniero, sono stati perseguiti penalmente. Gli operatori sono stati infine assolti, ma la vicenda ha chiarito le implicazioni dell'attività investigativa nello spazio virtuale (cyber spazio).

Il 2° Protocollo addizionale alla Convenzione di Budapest, aperto alla firma, affronta questo problema attraverso strumenti di consenso preventivo, basati sulla stabilizzazione delle squadre miste. Tuttavia, la soluzione non è ancora adeguata alla velocità e alla moltiplicazione delle operazioni da ricostruire.

Nell'ordinamento italiano non sorgono problemi diversi per il recente provvedimento (D.L. 9 agosto 2022 n. 115) che attribuisce alle agenzie di intelligence il potere di operare in modo difensivo anche con azioni offensive. La catena di responsabilità, tuttavia, non sembra considerare la necessaria velocità delle azioni difensive, anche negli spazi di sovranità altrui.

Il diritto internazionale pubblico fornisce strumenti chiari per consentire l'attività di verifica, nei limiti della legittima difesa, anche in relazione al principio di due diligence e a quelli aggiuntivi di necessità, proporzionalità e continenza. L'applicazione concreta di questi principi è ovviamente molto più discutibile.

Queste prime acquisizioni devono essere discusse in relazione ai recenti eventi di disinformazione. Il CCDCOE ha una grande esperienza e ha saputo valutare in modo approfondito quanto emerso dal conflitto ucraino, che ha visto le varie forme di condizionamento dell'opinione pubblica.

Vorremmo quindi poter accedere alle esperienze di Hybrid CoE e CCDCOE sui seguenti aspetti, anche dal punto di vista tecnico. Tutti gli argomenti si basano su ciò che è emerso nelle recenti esperienze e costituiscono un campo di discussione, così come di domande.

1.1. Quanto sono rilevanti i progressi nella costruzione di FN in grado di superare le difese di resilienza dei sistemi occidentali? Questi progressi rendono più difficile ricostruire l'origine del FN? In cosa consistono esattamente i passaggi che ne ricostruiscono l'origine? Quali sono i tempi di reazione utili per ricostruire i passaggi? In altri termini, questi passaggi costituiscono una traccia indelebile (come nel caso dei Block-Chain) o sono immediatamente o rapidamente modificabili dagli attori malintenzionati? Le informazioni

acquisite vengono trasformate dall'azione per ricostruire i passaggi durante la procedura di monitoraggio? In caso affermativo, sono possibili protocolli di verifica per garantire l'integrità delle informazioni a fini probatori? Quali sono le condizioni e le procedure per le operazioni di contrasto? Le LEA possono acquisire le informazioni utili per l'indagine attraverso l'accesso diretto a banche dati e server stranieri?

1.2. Sono stati accertati casi in cui la diffusione di FN è stata facilitata dall'uso dell'IA? Ad esempio, sono stati riscontrati casi in cui l'IA ha reso possibile adattare le informazioni false alla risposta (e.g.by introducendo automaticamente risposte che contrastano la probabilità di risultati volti a dimostrare la falsità – o l'adattamento della narrazione alla contro-narrativa)? In che modo esattamente l'IA cambia il modo in cui FN opera (oltre a GAN, NLP e simili)? L'AI rende la diffusione di FN più veloce ed efficace, oltre a quanto già realizzato con l'utilizzo dei BOT?

1.3. Metodi per valutare l'efficacia della disinformazione attraverso FN. È possibile accertare l'efficacia del FN nell'influenzare l'opinione pubblica? Vi sono casi di perseguimento effettivo dell'impiego di FN (al di là e al di là della semplice tutela del diritto dell'individuo alla personalità e all'integrità personale, come la sostituzione della persona o la diffamazione)? In caso affermativo, quali problemi giuridici sono stati affrontati?

2.1 In termini generali, in che modo la disinformazione viene rafforzata dall'IA? Sono state individuate campagne di disinformazione basate su tecniche diverse da quelle del FN? Su quali basi si può distinguere una campagna di questo tipo dalla formazione di un orientamento nell'opinione pubblica? Nella esperienza accumulata, quanto sono importanti gli effetti della conferma di pregiudizi o convinzioni preconcette degli utenti regolari? Quali lezioni sono state apprese dalle campagne di disinformazione non basate sull'uso di strumenti digitali (ad esempio, nel Sahel?) Come distinguere la disinformazione organizzata dalla diffusione di notizie avverse? In che modo l'IA facilita le campagne di disinformazione in un ambiente virtuale? Sono stati utilizzati casi concreti di strumenti di intelligenza artificiale per questo scopo? Nell'elaborazione delle notizie, nella loro diffusione, nell'immediato contrasto delle contro-narrazioni. In questi casi, quali sono i mezzi con cui è possibile identificare la presenza di agenti non umani? Il semplice uso di BOT è un'indicazione esaustiva del carattere non umano dell'interlocuzione? [5]

2.2 Sono stati sviluppati parametri per classificare una notizia come parte di una campagna di disinformazione? È possibile risalire all'origine della notizia una volta che è stata ripresa e diffusa autonomamente da soggetti umani?

2.3 Risalire la catena della disinformazione in questi casi comporta problemi simili a quelli esaminati al punto 1? Sono stati analizzati casi in cui queste tecniche sono state accompagnate da attacchi informatici tradizionali (penetrazione di sistemi informatici; utilizzo di informazioni reali, acquisite attraverso la penetrazione, ecc.)? In che modo gli strumenti di guerra cognitiva si combinano con altri strumenti di minaccia ibrida nelle operazioni contro le nostre società?

2.4 Vi sono casi in cui questo specifico tipo di campagna è stato oggetto di indagini penali? In caso affermativo, quali problemi sono stati affrontati? Ci sono state forme di cooperazione internazionale efficace con i paesi amici da cui è venuta parte dell'attacco?

2.5 La reazione all'attacco di disinformazione (nei limiti dell'IL e del diritto interno) può consistere in campagne di informazione (che per il paese a cui è soggetta sono considerate disinformazione) e quali sono i limiti? Possono essere attuati anche verso l'opinione pubblica interna e, se sì, con quali limiti e in che modo?

3.1 In generale, quale grado di affidabilità può essere effettivamente riconosciuto agli strumenti di attribuzione? Nel campo dell'IL, quale ruolo gioca nell'attribuzione il dato tecnico (ad esempio, la catena di passaggi ricostruiti nello spazio virtuale) rispetto a quelli circostanziali (*cui prodest*; linguaggio o carattere grafico; orientamento politico, ecc.)? Quali problemi tecnici sono stati affrontati nell'attribuzione? In

[5] Quale tecnologia può rispondere alle domande chiave su: - c'è una campagna in corso? – Da dove ha avuto origine? – Chi lo sta conducendo? – Quali potrebbero essere i suoi obiettivi?

particolare, l'AI può essere utilizzata come strumento investigativo a questo scopo? Con quali mezzi e con quale efficacia? Quali problemi giuridici sono sorti nella pratica e quali sono immaginabili?

4.1 Quali sono le linee di modifica del Manuale Tallinn 2 nei settori sopra discussi? Da quali constatazioni di fatto o da quali riflessioni giuridiche hanno avuto origine queste linee di modifica?

10 dicembre 2022
FVO