A cura di
Francesco Micozzi

TRANSIZIONE DIGITALE E IA

TRIS - Recupero³
Progetto finanziato dal MIMIT - Art. 148 L. 388/2000
- D.M. 6/5/2022 art. 5

TRANSIZIONE DIGITALE
E
INTELLIGENZA ARTIFICIALE

A CURA DI
FRANCESCO PAOLO MICOZZI

Il volume si inserisce nell'attività scientifica del Centre of Excellence Jean Monnet "BALDUS" (Building the Age of a Lawful and Sustainable Data-USe) dell'Università degli Studi di Perugia – EACEA 2021-2024 101047644 — BALDUS — ERASMUS-JMO-2021-HEI-TCH-RSCH

progetto TRIS - RECUPERO - finanziato dal Ministero dell'Impresa e del Made in Italy, Art. 148 L. 388/2000 - D.M. 6/5/2022 art. 5.

Indice

1. Cos'è l'intelligenza artificiale

La difficoltà di inquadrare chiaramente cosa sia "intelligenza artificiale", lo rende un concetto evanescente e, spesso, impalpabile. Sentiamo parlare di intelligenza artificiale negli ambienti più disparati e con riguardo ad oggetti che, tra loro, appaiono distanti e differenti. Ciò ha portato taluno a ritenere preferibile parlare di "intelligenze artificiali", proprio in considerazione dell'eterogeneità di oggetti a cui tale intelligenza artificiale viene riconosciuta. Questa difficoltà è, in parte, riferibile alla stessa difficoltà di inquadramento dei concetti alla base dell'intelligenza artificiale: cosa è l'intelligenza? Se ci ponessero di fronte a tale domanda, dimostreremmo – tentando di darvi risposta – la nostra difficoltà in tal senso. In prima battuta potremmo essere portati a definire l'intelligenza come la capacità delle persone risolvere specifici problemi sulla base delle informazioni acquisite nella fase di apprendimento (sia chetragga la sua origine in una componente innata a ciascuno qual è quella emotiva, o nell'esperienza personale o, ancora, nel nostro essere animali sociali). Probabilmente, però, penseremmo che l'intelligenza non si possa limitare al modo in cui applichiamo conoscenze precedentemente acquisite ma anche al modo in cui le acquisiamo. Potremmo, inoltre, ritenere che l'intelligenza non sia qualcosa necessariamente appannaggio degli esseri umani ma, potremmo riconoscere altre intelligenze in altri esseri viventi. In tal senso potremmo, quindi, definire l'intelligenza come la capacità di un essere vivente di comprendere, imparare, risolvere problemi, adattarsi e compiere scelte appropriate in base alle circostanze. Ma, anche questa definizione potrebbe non essere pienamente soddisfacente, considerando che – pur limitando la nostra indagine agli esseri umani – alcuni soggetti dimostrano un'intelligenza "superiore" rispetto ad altri limitatamente a specifici settori. Potremmo, quindi ritenere, come già suggerito dalla teoria delle intelligenze multiple di H. Gardner, che esista un'intelligenza logico-matematica, una linguistica, una spaziale, una musicale, una cinestetica, una interpersonale, una intrapersonale o una naturalistica.

Potremmo anche distinguere le manifestazioni dell'intelligenza – come suggerito da D. Kahneman – in pensiero veloce e pensiero lento. Il primo più "istintivo" e reattivo, il secondo più ponderato e riflessivo. Kahneman definisce pensiero veloce, in sostanza, quello automatico, intuitivo, rapido. È quell'elaborazione mentale immediata o "istintiva" che impieghiamo quotidianamente in attività "apparentemente" semplici come riconoscere un volto, reagire a un pericolo improvviso o interpretare espressioni emotive. Questo modo di pensare si basa su "scorciatoie" mentali e, per questo, può condurre a errori di giudizio. Il pensiero lento, invece, è – nella definizione di Kahneman – quello che richiede uno sforzo consapevole, è analitico e deliberato, e viene usato per compiti complessi che richiedono concentrazione e logica, come risolvere problemi matematici o fare valutazioni importanti. Essendo più lento e faticoso da usare, viene attivato solo quando il pensiero veloce non è sufficiente o risulta inefficace. Questi due sistemi di pensiero (quello lento e quello veloce) non sono indipendenti tra loro: collaborano costantemente. Tuttavia, il pensiero veloce che è maggiormente soggetto ad errori di valutazione può passare al pensiero veloce degli input (sotto forma di intuizioni o impressioni) errati che pregiudicano il risultato finale. Kahneman, pertanto, suggerisce di sviluppare consapevolezza sugli errori sistematici del pensiero veloce e di usare il pensiero lento per migliorare le nostre decisioni in contesti critici. Per Kahneman, quindi, l'intelligenza non è solo un'abilità razionale e logica, ma è strettamente collegata alla capacità di usare in modo efficace i due sistemi di pensiero.

Quando nel 1955, John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon, usano per la prima volta il concetto di "Intelligenza artificiale"[1] nel

[1] Si fa riferimento al documento "Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" (al quale seguirà, nell'estate del 1956 la storica conferenza di Dartmouth sull'intelligenza artificiale).

documento "Proposal for the Dartmouth Summer Research Project on Artificial Intelligence", il concetto di “intelligenza” al quale si pensava era sostanzialmente connesso a un sogno: quello di creare macchine in grado di simulare ogni aspetto dell'intelligenza umana. Nella definizione di intelligenza artificiale contenuta nel documento di presentazione della Conferenza di Dartmouth, quindi, abbiamo macchine create dall’uomo (artificiale) in grado di mostrare una qualche forma di intelligenza[2]. Gli obiettivi degli studi della Conferenza di Dartmouth erano quelli di studiare la possibilità, per le macchine di esibire comportamenti che sarebbero considerati intelligenti se fossero eseguiti da un essere umano (quali l’uso del linguaggio, la capacità di formare astrazioni e concetti, risolvere problemi e migliorarsi costantemente) e dimostrare che ogni aspetto dell'apprendimento e dell'intelligenza può essere descritto con precisione tale da poter essere simulato da una macchina.

2. A che punto siamo

A partire dal 30 novembre 2022, OpenAI, l'organizzazione statunitense all'avanguardia nella ricerca sull'IA, ha reso disponibile al grande pubblico ChatGPT, un avanzato modello linguistico capace di generare testo in modo “straordinariamente umano”. Questo evento ha catalizzato l'interesse globale verso l’IA, trasformandola da concetto di nicchia a fenomeno mainstream. A soli cinque giorni dal lancio, ChatGPT contava già un milione di utenti, mentre due mesi più tardi erano saliti a cento milioni. Il lanciodi ChatGPT ha immediatamente catturato l'attenzione globale: la capacità di ChatGPTdi sostenere conversazioni più o meno coerenti[3], di rispondere a domande complesse

[2] In futurologia si fa riferimento al concetto di “singolarità tecnologica”, con riferimento al punto di svolta ipotetico in cui l'intelligenza artificiale e la tecnologia avanzata superano l'intelligenza umana, causando cambiamenti così rapidi e profondi da risultare incontrollabili e imprevedibili per l'umanità. Si veda al riguardo VINGE, *Technological singularity*, in VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, 1993, pp. 30–31.

[3] Si evidenzia che un sistema di intelligenza artificiale generativa come ChatGPT ha capacità di elaborazione del linguaggio naturale basata su pattern statistici ma non ha capacità di comprensione semantica profonda.

e di fornire assistenza in una varietà di contesti ha aperto nuove prospettive sull'interazione tra uomo e macchina (o meglio, tra uomo e intelligenza artificiale). Questa interazione ha generato stupore e curiosità, portando molti a riflettere su come l'IA influenzi diversi aspetti della società. L'impatto di ChatGPT è stato talmente profondo che molti hanno iniziato a identificare l'intera IA con questo strumento: ChatGPT è diventato il volto dell'IA per milioni di persone, al punto che molti lo percepiscono come la rappresentazione dell'intelligenza artificiale e che fanno coincidere l'avvento dell'era dell'intelligenza artificiale proprio con il lancio di ChatGPT. Tuttavia, in modo apparentemente silenzioso, altri sistemi di intelligenza artificiale interagivano già col grande pubblico. Si pensi al sistema di "autocompletamento" di Google, in base al quale quando iniziamo a digitale una ricerca, Google suggerisce automaticamente termini e frasi, prevedendo ciò che stiamo cercando grazie ad algoritmi di apprendimento automatico. Si pensi, ancora ai sistemi di raccomandazione di Netflix e Spotify (che analizzando le abitudini di visione o ascolto degli utenti suggeriscono film, serie TV o brani musicali in linea con i loro gusti), agli assistenti virtuali come Siri e Alexa (che utilizzano l'IA per comprendere comandi vocali e fornire risposte o eseguire azioni specifiche), ai filtri antispam nelle email (offerti dai provider di posta elettronica che impiegano algoritmi di IA per identificare e filtrare messaggi indesiderati o potenzialmente pericolosi), alla pubblicità personalizzata sui social media (è con l'IA che si analizzano le interazioni degli utenti e si mostrano loro annunci mirati), e, ancora traduttori automatici, riconoscimento facciale negli smartphone per sbloccare il dispositivo, sistemi di navigazione che ci forniscono percorsi ottimali e tempi di arrivo stimati, strumenti di controllo ortografico e grammaticale nei programmi di videoscrittura e suggerimenti per gli acquisti sui market online.

ChatGPT è, di fatto, solo uno dei molti approcci all'intelligenza artificiale, basato su modelli linguistici che predicono sequenze di parole piuttosto che "comprendere"

realmente. Questa popolarità senza precedenti e la percezione pubblica evidenziano il bisogno di maggiore educazione sull'IA: cosa può e cosa non può fare, quali sono i suoi limiti e quali le sue potenzialità reali e, naturalmente, quali sono i rischi diretti o indiretti che possono discendere dall'uso di un sistema di intelligenza artificiale. La consapevolezza crescente ha dato vita a dibattiti sulle implicazioni etiche dell'IA. Questioni come la privacy, la sicurezza dei dati e l'impatto sul mercato del lavoro sono diventate centrali nelle discussioni pubbliche.

L'intelligenza artificiale moderna rappresenta uno dei più affascinanti esempi di come il progresso scientifico e tecnologico non proceda per compartimenti stagni, ma attraverso una complessa rete di interconnessioni tra diverse discipline e correnti di pensiero. Quando oggi osserviamo un sistema di IA in azione, stiamo in realtà guardando il risultato di secoli di evoluzione del pensiero umano, dove filosofia, matematica, biologia, psicologia e informatica si sono intrecciate in modo inestricabile.

L'intelligenza artificiale moderna rappresenta il culmine di un viaggio intellettuale millenario, le cui radici affondano nell'antica Grecia. **Aristotele**, nel IV secolo a.C., pose le prime pietre di questo edificio attraverso la sua logica formale e il sistema dei sillogismi. La sua intuizione che il ragionamento potesse essere formalizzato in regole precise ha aperto la strada a tutti gli sviluppi successivi nel campo della logica computazionale. Durante l'età dell'oro islamica, Mufiammad ibn Mūsā **al-Khwārizmī** rivoluzionò il pensiero matematico introducendo il concetto di algoritmo - termine che deriva proprio dal suo nome. I suoi contributi all'algebra e all'introduzione del sistema numerico indo-arabico in Occidente crearono le basi per la matematica computazionale. Questo patrimonio fu poi arricchito dal **Fibonacci** nel XIII secolo, che non solo diffuse ulteriormente il sistema numerico posizionale in Europa, ma introdusse anche sequenze matematiche che oggi trovano applicazione negli algoritmi di machine learning. **Ramon Llull**, filosofo catalano del XIII secolo, fece un passo

ulteriore immaginando una macchina logica capace di combinare concetti fondamentali per generare nuove verità. Il suo "Ars Magna" può essere considerato uno dei primi tentativi di meccanizzare il pensiero logico. Secoli dopo, **Cartesio** avrebbe ripreso questa sfida da una prospettiva filosofica diversa, introducendo il dualismo mente-corpo e ponendo questioni sulla natura della coscienza che ancora oggi influenzano il dibattito sull'IA. Il XVII secolo vide l'emergere di due figure cruciali: Blaise Pascal e Gottfried Leibniz. **Pascal** costruì la prima calcolatrice meccanica funzionante, mentre **Leibniz** non solo perfezionò questo strumento, ma concepì anche l'idea di un linguaggio universale del ragionamento, il "*calculus ratiocinator*". La visione di Leibniz di poter ridurre il pensiero a calcolo matematico anticipa in modo sorprendente molti aspetti dell'IA moderna. L'Ottocento segnò una svolta decisiva con George **Boole**, che sviluppò l'algebra booleana, fornendo il fondamento matematico per la logica digitale. Contemporaneamente, Charles **Babbage** progettò la sua Macchina Analitica, il primo computer programmabile della storia, mentre Ada **Lovelace** intuì le potenzialità di questa macchina oltre il puro calcolo numerico, diventando la prima programmatrice della storia. Il XX secolo ha visto l'emergere di due figure che hanno definitivamente gettato le basi dell'IA moderna: **Alan Turing** e John von Neumann. Turing non solo definì il concetto teorico di computabilità attraverso la sua macchina ideale, ma pose anche le basi per la valutazione dell'intelligenza artificiale attraverso il famoso "Test di Turing". Il suo lavoro sulla decrittazione durante la Seconda Guerra Mondiale dimostrò il potere pratico del calcolo automatico. **Von Neumann**, d'altra parte, sviluppò l'architettura di base dei computer moderni e contribuì alla teoria degli automi cellulari. La sua visione di macchine auto-replicanti e la sua comprensione del parallelismo tra sistemi biologici e computazionali hanno influenzato profondamente lo sviluppo dell'IA.

Questa lunga catena di innovazioni intellettuali dimostra come l'IA non sia emersa improvvisamente e dal nulla, ma sia il risultato di un processo cumulativo di intuizioni

e scoperte. Ogni pensatore ha aggiunto un tassello essenziale: dalla logica formale di Aristotele agli algoritmi di al-Khwārizmī, dalla visione meccanicistica di Cartesio alla logica booleana, fino alle fondamentali intuizioni di Turing e von Neumann.

Questa ricca eredità intellettuale ha permesso alle menti della Dartmouth Conference di immaginare la possibilità concreta di macchine intelligenti, basandosi su secoli di sviluppo del pensiero logico, matematico e filosofico.

3. Cosa fa (e cosa non fa) una IA Generativa, come ChatGPT?

Il grande pubblico, come abbiamo visto, ha preso un primo contatto con l'Intelligenza artificiale a novembre 2022 con ChatGPT. Quest'ultimo è un modello di intelligenza artificiale sviluppato da OpenAI, progettato per interagire con gli utenti in modo conversazionale. Il nome "ChatGPT" deriva dalla combinazione di "Chat", che indica la sua funzione di chatbot per la comunicazione interattiva, e "GPT", acronimo di Generative Pre-trained Transformer. Il funzionamento di ChatGPT si basa su un processo di addestramento in due fasi: pre-addestramento e fine-tuning. Durante il pre-addestramento, il modello viene esposto a una grande quantità di dati testuali provenienti da fonti disparate (non è possibile conoscere quali fonti siano state impiegate), apprendendo le strutture grammaticali, il vocabolario e le informazioni generali. Successivamente, nel fine-tuning, il modello viene raffinato su dataset più piccoli e specifici, spesso con supervisione umana, per migliorare le sue prestazioni in compiti particolari e per renderlo più adeguato alle interazioni con gli utenti.

I sistemi di Intelligenza Artificiale (IA) generativa rappresentano una delle innovazioni più significative nell'ambito dell'IA moderna, frutto di decenni di progresso nel campo dell'apprendimento automatico e delle reti neurali profonde. Essi sono in grado di creare contenuti originali, come testi, immagini, musica e video, sfruttando modelli

avanzati di apprendimento automatico. Questi sistemi nascono dall'esigenza di creare macchine in grado di generare contenuti originali, come testi, immagini, musica o video, che siano indistinguibili da quelli prodotti da un essere umano.

Il processo di addestramento di una IA generativa inizia con la raccolta di enormi dataset[4] pertinenti al tipo di contenuto che si desidera generare. Ad esempio, per un modello di linguaggio naturale finalizzato alla generazione di testo, si utilizzano miliardi di testi tratti da libri, articoli e siti web. Questi dati vengono poi utilizzati per addestrare reti neurali profonde[5], in particolare modelli "basati su *transformer*"[6], che, grazie al meccanismo di *self-attention*[7], sono in grado di mettere in relazione direttamente ogni parola di una sequenza con tutte le altre, indipendentemente dalla distanza tra le stesse[8]. Durante l'addestramento, il modello impara a prevedere l'elemento successivo in una sequenza, dato il contesto fornito dagli elementi precedenti. Questo processo, noto come modellazione linguistica, permette alla IA di apprendere le regole grammaticali, le strutture sintattiche e le semantiche del linguaggio[9].

[4] Un dataset è un insieme strutturato di dati organizzati in modo tale da poter essere utilizzati per analisi, elaborazioni o addestramento di modelli di AI.

[5] Le reti neurali profonde (o *deep neural networks*, DNN) sono un tipo di modello matematico ispirato al funzionamento del cervello umano, utilizzato nell'intelligenza artificiale per elaborare grandi quantità di dati e risolvere problemi complessi. Si chiamano "profonde" perché sono composte da molti strati (layer) di neuroni artificiali (unità computazionali o nodi), organizzati in una struttura gerarchica.

[6] BENGESI ET AL., *Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers*, IEEE Access, 2024, 69812–69837.

[7] VASWANI ET AL., *Attention is All you Need*, 2017.

[8] Le dipendenze a lungo raggio (in inglese, *long-range dependencies*) sono relazioni o connessioni che si verificano tra elementi distanti all'interno di una sequenza o di una struttura. Nel linguaggio, le dipendenze a lungo raggio emergono quando la comprensione di una parola o di una frase dipende da elementi lontani nella sequenza. Ad esempio: "Il gatto che il bambino ha accarezzato ieri sta sulla poltrona." Per determinare che il verbo "sta" si riferisce al soggetto "il gatto", bisogna considerare una connessione attraverso diversi elementi intermedi ("che il bambino ha accarezzato ieri").

[9] L'addestramento avviene attraverso l'ottimizzazione di una funzione di perdita, che misura la differenza tra le previsioni del modello e i dati reali, e viene minimizzata utilizzando algoritmi di *backpropagation* e ottimizzatori come Adam o RMSprop. I meccanismi che permettono al modello di funzionare efficacemente includono l'attenzione (*attention mechanism*), che consente al modello di focalizzarsi sulle parti rilevanti dell'input durante la generazione dell'output. Questo è fondamentale per mantenere la coerenza e la coesione nel contenuto generato. Inoltre, vengono utilizzate tecniche di pre-addestramento e fine-tuning: il

Nonostante la diversità tra i vari sistemi di IA, esistono caratteristiche comuni che definiscono le potenzialità dei sistemi generativi.

I sistemi di IA generativa operano con una **velocità** nell'elaborazione e nell'analisi di dati su **larga scala**, che, in molti ambiti, supera in modo evidente quella umana. Questa rapidità è dovuta all'utilizzo di algoritmi ottimizzati e infrastrutture hardware potenti, come unità di elaborazione grafica (GPU) e *tensor processing units* (TPU). La capacità di processare grandi volumi di dati in tempi ridotti permette a questi sistemi di generare output (testi, audio, video etc) in modo quasi istantaneo. Ad esempio, un modello di linguaggio generativo può produrre un articolo completo in pochi secondi, accelerando processi editoriali e creativi. I sistemi di IA generativa sono progettati, inoltre, per gestire simultaneamente molteplici attività (**multitasking**). Grazie alle reti neurali profonde e agli algoritmi di apprendimento parallelo, possono eseguire diverse operazioni contemporaneamente senza compromettere la qualità del risultato. Questa caratteristica è particolarmente utile in applicazioni complesse dove è necessario integrare diverse fonti di informazione e generare output multidimensionali. Ad esempio, un assistente virtuale può comprendere comandi vocali, analizzare il contesto e fornire risposte pertinenti, tutto in tempo reale. Una delle potenzialità più significative dei sistemi di IA generativa è la loro abilità del mettere tra loro in **relazione, confrontare e analizzare** dati. Con tecniche di machine

modello viene prima addestrato su un ampio corpus di dati generici e successivamente raffinato su dataset specifici per adattarlo a compiti particolari. Gli algoritmi alla base del funzionamento delle IA generative sono spesso basati su architetture di reti neurali ricorrenti (RNN) o trasformatori (*Transformers*). I trasformatori, in particolare, hanno rivoluzionato il campo grazie alla loro capacità di gestire parallelamente l'elaborazione delle sequenze, rendendo l'addestramento più efficiente. Un esempio emblematico è il modello GPT (*Generative Pre-trained Transformer*), che utilizza strati di attenzione per modellare le relazioni tra le parole in un testo. Per generare contenuti, il modello utilizza metodi di campionamento come il campionamento casuale, la ricerca del fascio (*beam search*) o il campionamento top-k/top-p, che gli permettono di selezionare le opzioni più probabili in base alla distribuzione appresa durante l'addestramento.

learning, come il deep learning[10] e l'apprendimento per rinforzo[11], questi sistemi possono identificare pattern nascosti[12] e correlazioni in dataset di grandi dimensioni. **L'automazione** è al centro dei sistemi di IA generativa. La capacità di svolgere autonomamente compiti ripetitivi o complessi è, infatti, una delle caratteristiche della AI generativa.

Di converso, però, nonostante le avanzate capacità di elaborazione, i sistemi di Intelligenza Artificiale generativa attuali mancano di una vera comprensione del contesto e delle sfumature culturali ed emotive. Operando principalmente attraverso modelli statistici e calcoli probabilistici, generano output basati sulla probabilità che una certa sequenza di parole sia appropriata, senza una reale comprensione semantica. Questa limitazione può portare alla produzione di contenuti che, sebbene grammaticalmente corretti, mancano di profondità semantica o possono risultare incoerenti e privi di senso in determinati contesti. La mancanza di sensibilità culturale ed emotiva significa anche che l'IA generativa può non cogliere sfumature importanti, portando a risultati che non rispecchiano accuratamente le intenzioni o le aspettative

[10] Il deep learning è un ramo dell'intelligenza artificiale che permette ai computer di imparare e prendere decisioni simili a quelle umane, usando strutture chiamate reti neurali profonde. Si chiama "deep" (profondo) perché queste reti hanno molti strati, ognuno dei quali elabora i dati in modo più complesso. Si potrebbe rappresentare una rete neurale come una serie di filtri: ogni strato della rete acquisisce informazioni in ingresso da un nodo, le elabora e le passa a un ulteriore nodo, finché non si giunge al risultato definitivo.

[11] L'apprendimento per rinforzo, noto anche come *reinforcement learning*, è una tecnica di machine learning ispirata al modo in cui gli esseri umani e gli animali apprendono attraverso l'esperienza. L'idea di base è che un agente, cioè un programma o un algoritmo, interagisce con un ambiente e impara a prendere decisioni per raggiungere un obiettivo. Questo apprendimento avviene grazie a un sistema di ricompense e penalità che guida l'agente verso comportamenti sempre più efficaci. Immaginiamo un robot che deve imparare a camminare. All'inizio, non sa come muoversi: prova a spostare un piede, cade, si rialza, e prova di nuovo. Ogni volta che riesce a mantenere l'equilibrio, riceve una ricompensa, mentre ogni caduta gli fornisce un segnale negativo. Il robot registra questi risultati e, attraverso molti tentativi ed errori, capisce quali movimenti lo avvicinano al suo obiettivo. Questo processo di apprendimento non è guidato da istruzioni dirette, ma dalla continua interazione con l'ambiente, dove l'agente cerca di massimizzare le ricompense nel lungo termine.

[12] I pattern nascosti sono schemi o regolarità presenti in un insieme di dati o in un sistema, ma non immediatamente visibili o riconoscibili a prima vista. Questi pattern non sono evidenti perché possono essere mascherati dalla complessità, dal rumore, o dalla grande quantità di informazioni, e spesso richiedono tecniche avanzate per essere individuati, come l'analisi statistica, il machine learning o l'intelligenza artificiale.

umane. Il fenomeno delle **allucinazioni**[13] dell'Intelligenza Artificiale è un esempio concreto di queste limitazioni. Si verifica quando l'IA genera informazioni inesatte o completamente inventate, pur mantenendo una struttura linguistica plausibile. Questo avviene perché il modello non ha la capacità di verificare la veridicità delle informazioni che produce; si basa esclusivamente sulle correlazioni apprese dai dati di addestramento, senza una comprensione reale dei concetti. Questo ci porta a comprendere che l'output di una IA generativa non è frutto di una ricerca delle informazioni in un database o dell'uso di un motore di ricerca. Si tratta, invero, di calcoli di probabilità statistica tra i miliardi di numeri di un *large language model* (LLM). Ciò significa che potremmo definire allucinazione qualsiasi output di un LLM, tuttavia, "la chiamiamo così solo quando ci accorgiamo che la risposta è sbagliata". Il problema è che i LLM sono così bravi in quello che fanno che ciò che inventano è giusto (o, meglio, corrispondente alla risposta corretta) nella maggior parte dei casi. Comprendiamo, quindi, che è difficile fidarsi del risultato di un LLM: il testo generato viene scritto bene ma c'è una possibilità che la risposta non sia quella "corretta". Per questa ragione non ci si può fidare ciecamente delle risposte. È importante, quindi, che facciamo domande in ambiti nei quali siamo sufficientemente esperti da comprendere se la risposta sia corretta o sia, invece, un'"allucinazione".

4. Quanti tipi di IA esistono?

Oltre ai sistemi di IA analoghi a ChatGPT (LLM) esistono ulteriori tipologie di intelligenza artificiale. Le reti neurali convoluzionali (**CNN**), ad esempio, sono quelle impiegate nell'elaborazione di immagini e video. Grazie alle CNN sono possibili le soluzioni di riconoscimento facciale, diagnostica medica su immagini radiologiche, l'analisi di segnali e ostacoli stradali nelle auto a guida assistita, il monitoraggio delle

[13] DOUGLAS HEAVEN, *Why does AI hallucinate?*, MIT Technology Review, 2024.

colture e così via. Esistono, poi, le reti neurali ricorrenti (**RNN**) che sono impiegate per l'elaborazione di dati sequenziali nel riconoscimento vocale (si pensi agli assistenti domestici come Alexa o Siri o ai programmi di trascrizione automatica del parlato) o per le traduzioni automatiche o per l'analisi e le previsioni che si basano su dati acquisiti nel tempo (come nel caso delle previsioni finanziarie o meteorologiche. I sistemi di IA con apprendimento per rinforzo (**RL**), invece, aiutano le macchine a prendere decisioni basate sull'interazione con l'ambiente, basandosi sulla logica dei premi e delle punizioni (che insegnano alla macchina ad evitare di ripetere determinati errori o a perseguire in altri determinati ambiti). I sistemi di apprendimento con rinforzo vengono impiegati nell'uso della IA nei videogiochi, nell'addestramento di robot o nell'efficientamento di energia o traffico. Esistono, poi, i cosiddetti "**sistemi esperti**", ossia programmi che usano conoscenze codificate manualmente da esperti umani, e vengono usati in ambito sanitario per diagnosticare malattie, in ambito giuridico per analizzare norme o contratti, in ambito aziendale per il supporto nelle delle decisioni strategiche e così via. Le "**reti generative avversarie**" (GAN), inoltre, sono sistemi di intelligenza artificiale costituiti da due reti neurali (il "generatore" e il "discriminatore") che lavorano in competizione tra loro per generare dati realistici. In sostanza mentre il Generatore ha il compito di creare nuovi dati (come, ad esempio, immagini o suoni), cercando di renderli il più possibile simili ai dati reali, il Discriminatore agisce come un critico (riceve come input sia dati reali che dati creati dal Generatore e cerca di distinguere tra i due). Il Generatore tenta di "ingannare" il Discriminatore creando dati sempre più realistici, mentre il Discriminatore diventa sempre più abile nel distinguere i dati reali da quelli falsi. Questo processo di competizione porta il GAN a migliorarsi continuamente, fino a quando il Generatore è in grado di produrre dati talmente realistici da essere quasi indistinguibili da quelli reali da parte del Discriminatore. L'ambito applicativo dei sistemi GAN è quello della

generazione di immagini o nel miglioramento delle immagini (tipo restauro o aumento della risoluzione delle immagini).

5. Cosa dobbiamo aspettarci da qui ai prossimi anni

È difficile prevedere con precisione come evolveranno i sistemi di intelligenza artificiale nei prossimi anni, ma è probabile che assisteremo a diversi sviluppi significativi. Certo, difficilmente vedremo, da qui a pochi anni, l'avvento dell'AGI (Artificial General Intelligence), ossia quei sistemi di intelligenza artificiale – che al momento rappresentano unicamente un obiettivo ambizioso – capaci di comprendere, apprendere e applicare conoscenze in modo generale e avanzato, allo stesso modo in cui farebbe un essere umano. L'AGI dovrebbe avere la capacità di affrontare diversi problemi in differenti domini, senza necessità di programmazione specifica per ciascuno. Questi sistemi, tuttavia, non esistono ancora né può dirsi se o quando ci si potrà arrivare.

Ciò nonostante, nei prossimi anni assisteremo a una maggiore integrazione dell'intelligenza artificiale nelle nostre vite quotidiane, in settori come sanità, educazione, finanze e trasporti. È inoltre probabile che nei prossimi anni assisteremo a progressi nell'ambito dell'apprendimento automatico e nelle reti neurali profonde, che renderanno i sistemi di IA più efficienti e capaci di affrontare problemi complessi. L'IA generativa, come i modelli di linguaggio avanzati, potrebbe diventare ancora più sofisticata, migliorando la capacità di generare contenuti. Gli sforzi attuali nell'orientare i sistemi di IA verso una maggiore trasparenza, per garantire che le elaborazioni delle IA siano comprensibili e "giuste", porteranno a un miglioramento in questi ambiti, così come negli ambiti della sicurezza e dell'etica per prevenire abusi e a proteggere la privacy degli individui. È sempre difficile ipotizzare quale sarà il futuro

dell'IA, ma è sicuro che non si tratta di una "moda passeggera" ma di un ambito scientifico che troverà sempre maggiore applicazione nelle nostre vite quotidiane.

6. Le norme di regolamentazione delle AI

La regolamentazione dell'Intelligenza Artificiale (IA) è diventata un tema cruciale per garantire che lo sviluppo e l'implementazione di queste tecnologie avvengano in modo responsabile e in conformità con i diritti fondamentali. In ambito giuridico, i temi chiave riguardano la trasparenza e la spiegabilità, la protezione dei dati personali, l'equità e la non discriminazione, la responsabilità e, infine, le certificazioni.

Uno dei concetti fondamentali è quello dell'Intelligenza artificiale spiegabile (XAI). Considerando il ruolo sempre più pervasivo dell'IA nell'era digitale moderna nel supportare molte delle nostre scelte—dall'individuazione dell'hotel più adatto alle nostre esigenze alla scelta del volo più conveniente per raggiungere una destinazione specifica—è importante comprendere, di fronte all'aumento dell'utilizzo di algoritmi complessi e spesso opachi, come e quanto possiamo fare affidamento sulle decisioni prese dall'IA. Per rispondere a questa domanda, entrano in gioco le procedure dell'IA spiegabile (XAI).

L'IA spiegabile si riferisce a un insieme di metodi e processi che rendono comprensibili agli esseri umani le decisioni e il funzionamento degli algoritmi di IA. La XAI descrive un modello di IA, ne anticipa gli effetti e individua potenziali bias, contribuendo a caratterizzare l'accuratezza, l'equità, la trasparenza e i risultati nei processi decisionali basati sull'IA. L'obiettivo principale è fornire trasparenza e comprensibilità, permettendo agli utenti di capire come e perché una decisione è stata presa, verificare l'accuratezza e l'equità del processo decisionale e poter così fare affidamento sui risultati offerti dall'IA, sapendo che possono ottenere spiegazioni significative.

Le aziende utilizzano sempre più algoritmi per prendere decisioni che hanno impatti significativi sulle persone, o, se vogliamo, sui consumatori. Ad esempio:

- Piattaforme digitali come Amazon, Google e Facebook personalizzano ciò che gli utenti vedono, influenzando le informazioni e i prodotti a cui hanno accesso.
- Servizi finanziari utilizzano algoritmi per approvare prestiti o stabilire limiti di credito. Un esempio è quando una banca decide se concedere un mutuo basandosi su un algoritmo che analizza la storia creditizia di un individuo.
- Tecnologie assistive alla guida aiutano nelle decisioni di sterzata e frenata quando siamo al volante.

Questi algoritmi spesso funzionano come "scatole nere", ossia strumenti i cui meccanismi di funzionamento sono insondabili a causa della loro complessità. Il fatto che alcuni sistemi di elaborazione rappresentino delle “black box” deriva dall'elevato numero di nodi computazionali (immaginiamoli come le sinapsi di un cervello digitale): i sistemi di IA sono in grado di elaborare e correlare molte più informazioni di quante un essere umano possa esaminare, a posteriori.

Anche supponendo che sia possibile risalire, a ritroso, al procedimento di elaborazione (che, umanizzando l'IA, potremmo chiamare "ragionamento") seguito dalla macchina per giungere alla risposta partendo dal prompt, sarebbero necessari più anni di quanti ne contenga un'intera vita umana. Senza il concetto di “spiegabilità”, però, è difficile identificare e correggere i “pregiudizi” eventualmente presenti negli algoritmi. Ad esempio, se un algoritmo utilizzato per selezionare candidati per un lavoro tende a favorire un genere o un'etnia specifica a causa dei dati di addestramento, senza una spiegazione è complicato individuare e correggere questo bias.

La comprensione di come un sistema IA raggiunge un risultato specifico offre numerosi vantaggi. Permette agli sviluppatori di assicurarsi che il sistema funzioni correttamente, aiuta a soddisfare standard normativi e consente a chi è influenzato da una decisione di contestarla o modificarla. Ad esempio, se un cliente si vede negare un prestito, dovrebbe poter capire quali fattori hanno contribuito a quella decisione per poter eventualmente fornire ulteriori informazioni o correggere errori nei dati.

Per trasformare l'IA in XAI, vengono utilizzate tecniche specifiche che includono:

- Accuratezza delle previsioni: verificata attraverso simulazioni e confronti con i dati di addestramento per assicurarsi che l'algoritmo fornisca risultati corretti. Ad esempio, un modello predittivo nel settore sanitario deve essere accurato nel diagnosticare una malattia per essere utile.
- Tracciabilità: ottenuta limitando le decisioni a regole e funzioni definite, permettendo di seguire il percorso che ha portato a una determinata decisione. Ad esempio, in un sistema di raccomandazione di prodotti, tracciare perché un certo articolo è stato suggerito a un utente.
- Comprensione delle decisioni: coinvolge l'aspetto umano, educando il team a capire come e perché l'IA prende determinate decisioni, facilitando così la comunicazione con gli utenti finali.

La differenza tra interpretabilità e spiegabilità nell'IA risiede nel fatto che l'interpretabilità misura quanto un osservatore può comprendere il motivo di una decisione, mentre la spiegabilità analizza più a fondo come l'IA è giunta a quel risultato. L'IA spiegabile si concentra sui risultati dopo che sono stati calcolati, fornendo una retrospettiva sulle decisioni prese. L'IA responsabile, invece, riguarda la pianificazione preventiva per rendere l'algoritmo affidabile prima del calcolo dei risultati, incorporando principi etici e normativi nel design del sistema. Entrambe

possono collaborare per migliorare l'IA, garantendo che sia sia affidabile che comprensibile.

Una delle soluzioni per rendere maggiormente affidabili i risultati ottenuti tramite sistemi di IA è introdurre obblighi normativi e principi che stimolino la XAI. Implementare l'IA spiegabile porta numerosi vantaggi:

- Migliora la fiducia degli utenti: le persone sono più propense a utilizzare servizi che percepiscono come trasparenti e comprensibili.
- Riduce i bias: le spiegazioni aiutano a identificare e correggere pregiudizi negli algoritmi, promuovendo l'equità.
- Supporta la conformità normativa: aiuta a soddisfare requisiti legali come quelli del GDPR, che richiedono trasparenza nelle decisioni automatizzate.
- Migliora i processi interni: le organizzazioni possono ottimizzare i loro algoritmi comprendendo meglio come funzionano, aumentando l'efficienza e l'efficacia.

Ad esempio, una banca che utilizza un algoritmo di valutazione del credito spiegabile può non solo garantire che i prestiti siano concessi in modo equo, ma anche identificare opportunità per offrire nuovi prodotti finanziari basati su una comprensione più profonda delle esigenze dei clienti.

La promozione dell'Intelligenza artificiale spiegabile, quindi, è essenziale per assicurare che l'IA operi in modo trasparente, equo e responsabile. Attraverso l'adozione di obblighi normativi e principi etici, possiamo costruire sistemi di IA che non solo siano potenti e innovativi, ma anche degni della fiducia degli utenti e allineati con i valori fondamentali della società.

7. Breve introduzione all'AI Act

Abbiamo già avuto modo di osservare come l'intelligenza artificiale sia una delle tecnologie più rivoluzionarie del nostro tempo, capace di trasformare settori chiave come la sanità, i trasporti, l'istruzione e l'industria. Tuttavia, il suo potenziale non è privo di rischi: questioni etiche, la protezione dei dati, la sicurezza e il rispetto dei diritti fondamentali sono diventati temi centrali nel dibattito pubblico e politico degli ultimi anni. In questo contesto, l'Unione Europea ha introdotto, con il Regolamento (UE) 2024/1689[14], meglio conosciuto come AI Act, un quadro normativo ambizioso e pionieristico che mira a disciplinare lo sviluppo, l'immissione sul mercato e l'uso dei sistemi di intelligenza artificiale all'interno degli Stati membri.

Il regolamento si pone come obiettivo di garantire che l'IA sia utilizzata in modo sicuro, equo e trasparente, promuovendo al contempo l'innovazione e la competitività. L'AI Act ha un approccio unico e bilanciato: da un lato protegge i cittadini e i consumatori da rischi inaccettabili, come la manipolazione psicologica o la discriminazione algoritmica; dall'altro, incentiva la crescita tecnologica, offrendo supporto alle piccole e medie imprese, alle start-up e ai centri di ricerca.

Il regolamento è suddiviso in 13 Capi e 13 Allegati, che delineano un quadro normativo completo per l'IA nell'UE. I Capi I e II definiscono l'obiettivo del regolamento, che è quello di migliorare il funzionamento del mercato interno e promuovere un'IA antropocentrica e affidabile, garantendo un elevato livello di protezione della salute, della sicurezza e dei diritti fondamentali. Al Capo III si tratta della classificazione dei sistemi di IA. Si distingue tra sistemi vietati (ossia si proibiscono quei sistemi di IA che presentano rischi inaccettabili) e sistemi ad alto rischio (ossia che possono comportare rischi significativi per la salute, la sicurezza o i diritti fondamentali), imponendo requisiti rigorosi per la loro immissione sul mercato e utilizzo. Sono quindi disciplinati

[14] Adottato il 13 giugno 2024 e pubblicato nella Gazzetta Ufficiale dell'UE il 12 luglio 2024 - https://eur-lex.europa.eu/eli/reg/2024/1689/oj?locale=it

gli obblighi e responsabilità per gli operatori economici (fornitori, importatori, distributori e utilizzatori di sistemi di IA). Sono introdotte procedure per la valutazione della conformità dei sistemi di IA ad alto rischio e per la sorveglianza del mercato. È, inoltre, istituito un quadro di governance a livello dell'UE, promuovendo la cooperazione tra le autorità nazionali competenti e la Commissione europea per l'implementazione efficace del regolamento.

Ai sensi dell'art. 3 dell'AI Act, per "sistema di intelligenza artificiale" si intende, "*un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali*". La nozione, ai sensi del cons. 12, si incentra sulla capacità inferenziale, ossia sul processo di ottenimento degli output (risultato) da input (domanda), e include sia le tecniche che consentono l'inferenza attraverso approcci di apprendimento automatico (*machine learning-based systems*), che gli approcci basati sulla logica e sulla conoscenza codificata (*knowledge-based systems*).

Con "sistemi basati sull'apprendimento automatico" (o *machine learning-based systems*), si fa riferimento a tecniche di IA che usano dati e algoritmi per apprendere schemi e eseguire inferenze. I "sistemi basati sulla conoscenza" (*knlowledge-based systems*), invece, usano la logica, regole e conoscenza codificata "manualmente" per fare inferenze. Questi ultimi sistemi si basano su ontologie e ragionamento deduttivo.

Il funzionamento di un algoritmo a regole, che trova le proprie ascendenze nella Dartmouth Conference del 1956 (durante la quale, per la prima volta, si usa il concetto di "intelligenza artificiale"), presuppone la codificazione, in un programma, di regole del tipo ***if*** (condizione) ***then*** (azione), che consentono al sistema di eseguire l'azione ogni volta che la condizione si verifica, associando sia le condizioni che le azioni a una

probabilità, espressa in valori numerici tra 0 e 1, “che indica una soglia di attivazione che permette una maggiore flessibilità nel descrivere una situazione reale”.

Algoritmi di questo tipo consentono di ottenere output spiegabili, ma non scalabili né flessibili, potendo eseguire unicamente gli algoritmi definiti dai suoi creatori umani, e solo quando l’input si verifica proprio come previsto dagli stessi.

Di converso, le reti neurali “possono imparare da sole come risolvere un problema, senza richiedere che questo venga indicato tramite un algoritmo” e soltanto l’apprendimento inziale è normalmente supervisionato, attraverso soluzioni corrette generate da esseri umani.

Il fatto che l’AI Act abbia ricompreso nella definizione di sistemi di IA sia quelli basati su regole che quelli di machine e deep learning è funzionale allo scopo di assicurare che tutte le attività comprese nel ciclo di vita di detti sistemi di IA siano rispettose dei diritti umani – e, in particolare, l’eguaglianza e la non discriminazione, la dignità umana, la tutela dei dati personali, la salute, l’autonomia individuale e la protezione dell’ambiente, della democrazia e dello Stato di diritto, disegnando una IA antropocentrica, affidabile, sicura e trasparente, almeno con riferimento ai contenuti generati e alle interazioni con gli esseri umani, oltre che mantenuta sotto la supervisione umana.

L’AI Act adotta, a tal fine, un approccio basato sul rischio a cui i sistemi citati espongono, appunto, i valori tutelati, considerando che il rischio sussiste sia per entrambe le tipologie di IA (imponendo, di conseguenza, misure di protezione differenziate e graduate in ragione della gravità e della probabilità) che, in conseguenza del loro uso (immissione sul mercato o messa in servizio) si verifichino impatti negativi sui diritti umani, sulla democrazia e sullo Stato di diritto.

Restano esclusi dall’ambito applicativo dell’AI Act sia i sistemi connessi alla protezione degli interessi di sicurezza nazionale, e le attività di ricerca e sviluppo di sistemi di IA,

prima della loro messa a disposizione per l'uso o prova in condizioni reali che possano interferire coi valori tutelati.

L'AI Act definisce, in conseguenza del descritto approccio, il divieto di talune pratiche, in quanto espongono a rischi inaccettabili; stabilisce per i sistemi di IA ad alto rischio requisiti e obblighi in capo agli operatori, mentre vincola a requisiti di trasparenza determinati sistemi a basso rischio.

Tra le pratiche di IA vietate, sono comprese, ai sensi dell'art. 5:

a) l'immissione sul mercato, la messa in servizio o l'uso di sistemi di IA che usano tecniche subliminali, volutamente manipolative o ingannevoli;

b) l'immissione sul mercato, la messa in servizio o l'uso di sistemi di IA che sfruttano le vulnerabilità di persone fisiche singole o gruppi specifici di persone, dovute all'età, alla disabilità o a una specifica situazione sociale o economica, con l'obiettivo o l'effetto di distorcerne il comportamento in un modo che provochi o possa ragionevolmente provocare un danno significativo a tale persona o ad altre;

c) l'immissione sul mercato, la messa in servizio o l'uso di sistemi di valutazione o classificazione di persone o gruppi sulla base del loro comportamento sociale o di caratteristiche personali o della personalità, in cui il punteggio sociale (*social scoring*) così ottenuto possa determinare un trattamento pregiudizievole o sfavorevole in contesti sociali non collegati a quelli in cui i dati sono stati originariamente generati o raccolti, o comunque un trattamento pregiudizievole o sfavorevole ingiustificato o sproporzionato rispetto al comportamento sociale e alla sua gravità;

d) l'immissione sul mercato, la messa in servizio per tale finalità specifica o l'uso di sistemi di IA per effettuare valutazioni del rischio relative a persone fisiche al fine di valutare o prevedere il rischio che la persona commetta un reato, unicamente sulla base della sua profilazione o della valutazione dei tratti e delle caratteristiche della personalità, con l'esclusione dei sistemi di IA utilizzati a sostegno della valutazione

umana del coinvolgimento di una persona in un'attività criminosa, che si basino su fatti oggettivi e verificabili direttamente connessi a un'attività criminosa;

e) l'immissione sul mercato, la messa in servizio per tale finalità specifica, o l'uso di sistemi di IA che ampliano o creano banche date di riconoscimento facciale attraverso *scraping* non mirato di immagini da internet o da filmati di telecamere a circuito chiuso;

f) l'immissione sul mercato, la messa in servizio per tale finalità specifica o l'uso di sistemi di IA per inferire le emozioni di una persona fisica sul luogo di lavoro o negli istituti di istruzione, a meno che l'uso sia destinato a motivi medici o di sicurezza;

g) l'immissione sul mercato, la messa in servizio per tale finalità specifica o l'uso di sistemi di IA di sistemi di categorizzazione biometrica delle singole persone per trarne deduzioni o inferenze in merito a razza, opinioni politiche, appartenenza sindacale, convinzioni religiose o filosofiche, vita e orientamento sessuali, con eccezione dell'etichettatura e del filtraggio di set di dati biometrici acquisiti legalmente, e della categorizzazione di dati biometrici nelle attività di contrasto;

h) l'uso di sistemi di identificazione biometrica remota in tempo reale in spazi accessibili al pubblico a fini di contrasto, a meno che e nella misura in cui tale uso sia strettamente necessario per la ricerca specifica di persone scomparse, vittime di tratta o sfruttamento sessuale di esseri umani, ovvero per la prevenzione di una minaccia specifica, sostanziale e imminente per la vita o l'incolumità delle persone fisiche o di una minaccia reale e attuale o prevedibile di un attacco terroristico, o, infine, per la localizzazione o l'identificazione di una persona sospettata di aver commesso un reato,a fini di indagine, esercizio dell'azione penale o esecuzione della pena, per alcuni reati gravi dettagliati nell'allegato II, punibili con la privazione della libertà personale per una durata massima di quattro anni, sempre a condizione che sia stata condotta, prima dell'uso, la valutazione di impatto sui diritti fondamentali (FRIA) di cui all'art. 27

e il sistema di IA sia stato registrato nella banca dati dell'UE, a norma dell'art. 49. Il difetto di previa registrazione non inibisce l'uso di siffatti sistemi in situazioni di emergenza debitamente giustificate, a condizione che la registrazione sia completata senza indebito ritardo. Inoltre, l'utilizzo a fini di contrasto di siffatti sistemi, nelle descritte condizioni, è subordinato alla previsione, con norma nazionale, dei limiti e delle condizioni in cui ciò sia consentito e alla preventiva autorizzazione di una autorità giudiziaria o amministrativa indipendente. Tuttavia, in situazioni di urgenza debitamente giustificate, è possibile iniziare a usare il sistema senza autorizzazione, a condizione che l'autorizzazione sia richiesta senza indebito ritardo, al più tardi entro 24 ore; se l'autorizzazione fosse respinta, l'uso deve essere interrotto immediatamente, e tutti i dati, i risultati e l'output dell'uso devono essere immediatamente eliminati e cancellati. Inoltre, ogni uso deve essere notificato alle autorità nazionali di vigilanza del mercato e alle autorità nazionali per la protezione dei dati personali, le quali presentano alla Commissione europea relazioni annuali su tale uso, e, di conseguenza, la Commissione pubblica relazioni annuali.

I sistemi di IA sono classificati ad alto rischio, ai sensi dell'art. 6 dell'AI Act, se sono inseriti nel relativo allegato III, ovvero se sono soddisfatte due condizioni: a) che il sistema di IA sia destinato ad essere utilizzato come componente di sicurezza di un prodotto, o sia esso stesso un prodotto disciplinato dalla normativa di armonizzazione dell'Unione elencata nell'allegato I, e; b) che detto prodotto o sistema sia soggetto a una valutazione della conformità da parte di terzi ai fini dell'immissione o della messa in servizio, ai sensi della normativa di armonizzazione elencata nello stesso allegato I.

In deroga all'elenco di cui all'allegato III, non viene considerato ad alto rischio il sistema di IA che non presenta un rischio significativo di danno per la salute, la sicurezza o i diritti fondamentali delle persone fisiche, anche nel senso di non influenzare materialmente il risultato del processo decisionale. In questi casi, il fornitore deve documentarne la valutazione, prima della sua immissione sul mercato

o messa in servizio, ed è soggetto alla registrazione ai sensi dell'art. 49, oltre a dover mettere a disposizione delle autorità nazionali competenti, su loro richiesta, la documentazione relativa alla valutazione. Nell'allegato III – che può essere modificato dalla Commissione, sussistendone i presupposti – compaiono i sistemi utilizzati in taluni settori, tra i quali rilevano le infrastrutture critiche, che comprendono i sistemi di IA destinati a essere utilizzati come componenti di sicurezza nella gestione e nel funzionamento delle infrastrutture digitali critiche, del traffico stradale o nella fornitura di acqua, gas, riscaldamento o elettricità.

Il Regolamento sull'intelligenza artificiale europeo detta sia requisiti per l'immissione nel mercato e la messa in uso dei sistemi di IA ad alto rischio, che obblighi per i *deployer*, ossia per le persone fisiche e giuridiche, le autorità pubbliche, le agenzie o ogni altro organismo che utilizzi il sistema di IA sotto la propria autorità, eccezion fatta per l'uso nel corso di una attività personale non professionale.

8. La tutela dei consumatori nell'AI Act

L'AI Act, dopo aver dichiarato, nell'art. 1, che il suo scopo è quello di migliorare il funzionamento del mercato interno e promuovere la diffusione di un'intelligenza artificiale antropocentrica e affidabile, garantendo la tutela di diritti e interessi fondamentali ma promuovendo l'innovazione, precisa, nell'art. 2, che le norme dell'AI Act stesso lasciano "impregiudicate le norme stabilite da altri atti giuridici dell'Unione in materia di protezione dei consumatori e di sicurezza dei prodotti".

La premessa prende i passi dalla considerazione che sebbene sia chiaro a tutti quali possano essere i molteplici benefici e vantaggi offerti dai sistemi di intelligenza artificiale, non bisogna ignorare il fatto che da utilizzi impropri di tali nuovi e potenti strumenti possa derivare il rischio che siano implementate rinnovate e più potenti pratiche di manipolazione, sfruttamento e controllo sociale, rispetto a quelle sinora

conosciute. L'AI Act si preoccupa degli scenari in cui tali tecnologie siano sfruttate per la manipolazione e il controllo dei gusti, delle pulsioni e della libertà contrattuale dei consumatori. Per questo motivo, oltre a non alterare l'equilibrio normativo in materia di tutela e sicurezza dei consumatori, si propone di introdurre ulteriori norme a tutela dei consumatori in tutti quegli ambiti in cui siano impiegati sistemi di intelligenza artificiale. Attraverso l'IA, infatti, le più comuni tecniche manipolatorie usate per persuadere le persone ad adottare comportamenti indesiderati o per indurle con l'inganno a prendere decisioni in modo da sovvertirne e pregiudicarne l'autonomia, il processo decisionale e la libera scelta, possono assumere una dimensione diversa e più potente. Questo può comportare danni significativi con effetti negativi su salute fisica, psicologica o su interessi finanziari. Secondo l'AI Act, tali sistemi di IA, che impiegano componenti subliminali (quali stimoli audio, grafici e video non percepibili se non a livello subliminale) di matrice politica o commerciale, devono essere vietati e, quindi, messi al bando. L'autonomia, il processo decisionale e la libera scelta dei consumatori, infatti, potrebbe essere pregiudicati senza che gli stessi se ne rendano conto o che possano controllarle o resistervi. Questi effetti manipolatori potrebbero essere facilitati da interfacce cervello-computer o dalla realtà virtuale, in quanto maggiore sarebbe, con tali strumenti, la possibilità di controllo degli stimoli presentati alle persone. Inoltre, le tecnologie guidate da intelligenza artificiale potrebbero sfruttare le vulnerabilità o debolezze di particolari categorie di consumatori, maggiormente esposte alla circonvenzione in ragione della loro età, di una qualche disabilità o, ancora, in ragione di una situazione sociale o economica tale da rendere tali persone maggiormente esposte a tali rischi. Il divieto di tali pratiche, a mezzo di intelligenza artificiale, come visto, si pone come tutela complementare a quella stabilita dalle disposizioni della direttiva 2005/29/CE. In conseguenza di ciò, pertanto, le pratiche commerciali sleali che comportano danni economici o finanziari per i

consumatori sono vietate in ogni circostanza, indipendentemente dal fatto che siano attuate attraverso sistemi di intelligenza artificiale o in altro modo.

Il divieto di impiegare i sistemi di IA in pratiche di sfruttamento delle persone per distorcerne la libertà contrattuale, tuttavia, non riguarda anche le ipotesi in cui analoghe pratiche manipolative e di sfruttamento siano impiegate nel contesto di trattamenti medici (ad esempio in trattamenti di malattie mentali o nella riabilitazione fisica) qualora tali modalità siano disciplinate e regolamentate anche con norme specifiche del settore sanitario.

Allo stesso modo, l'AI Act non pone sullo stesso piano le tecniche manipolatorie che, tramite IA, incidano sull'autodeterminazione contrattuale delle persone ma rappresentino, invece, pratiche commerciali legittime, quali quelle in ambito pubblicitario. Si pensi, ad esempio, all'acquisto mirato di spazi pubblicitari sulla base dell'analisi preventiva effettuata tramite IA, oppure all'uso dell'analisi predittiva per identificare tendenze e migliorare le strategie pubblicitarie, o, ancora, agli strumenti di "*sentiment analysis*" (ossia al monitoraggio delle opinioni del pubblico verso un determinato prodotto o campagna pubblicitaria).

Il considerando 45 dell'AI Act ribadisce, quindi, il fatto che nessuna norma del regolamento deve incidere su pratiche già individuate come vietate da altre norme dell'Unione europea, tra le quali quelle in materia di protezione dei dati personali, di non discriminazione, della concorrenza e, ovviamente, a protezione dei consumatori. Ciò significa che se un'eventuale attività tramite sistemi di IA sia astrattamente ammessa da parte dell'AI Act ma vietata come pratica da altre norme europee, il divieto dovrebbe, comunque, svolgere i suoi effetti.

Come abbiamo già avuto modo di osservare, dalla crescente capacità dei sistemi di intelligenza artificiale di generare contenuti "sintetici"[15], indistinguibili da quelli reali, discendono significative sfide (oltre che criticità per la fiducia nel panorama informativo). Molto di ciò che appare reale, può non esserlo. I contenuti sintetici possono essere usati per manipolare i consumatori o generare disinformazione. Per tentare di limitare i rischi derivanti da un bombardamento di elementi informativi sintetici, il Regolamento AI Act, ai considerando 133 e 134, propone una maggiore trasparenza nei confronti dei consumatori. Per ottenere questa maggiore trasparenza, l'AI Act impone ai fornitori dei sistemi di IA in grado di generare contenuti sintetici di integrare soluzioni tecniche che consentano agli output di essere marcati in un formato leggibile meccanicamente e di essere rilevabili come generati o manipolati da un sistema di IA e non da esseri umani.

I fornitori di sistemi IA devono, in sostanza, integrare soluzioni tecniche affidabili che consentano di marcare i contenuti generati o manipolati artificialmente. Tra queste tecniche rientrano l'uso di filigrane digitali, metadati identificativi, metodi crittografici per attestare l'autenticità e la provenienza o registrazioni e impronte digitali dei contenuti. Le misure dovrebbero essere sufficientemente robuste e interoperabili, adattandosi all'evoluzione tecnologica e alle specificità dei diversi tipi di contenuti. L'obbligo si applica ai sistemi che modificano in modo sostanziale i dati di input,

[15] I contenuti sintetici e i dati sintetici sono informazioni generate artificialmente attraverso algoritmi e modelli matematici, spesso utilizzando tecniche di intelligenza artificiale. Sebbene i termini siano talvolta utilizzati in modo intercambiabile, esistono differenze significative tra i due concetti. Quando si parla di contenuti sintetici si fa riferimento a testi, immagini, audio o video creati da sistemi di intelligenza artificiale. Questi contenuti possono imitare o riprodurre fedelmente quelli generati da esseri umani, rendendo difficile distinguerli dagli originali. Un esempio emblematico sono i "deepfake", video o immagini in cui l'IA sovrappone il volto di una persona su un'altra, creando rappresentazioni altamente realistiche. L'uso di contenuti sintetici solleva questioni etiche e legali, specialmente quando vengono utilizzati per ingannare o manipolare l'opinione pubblica. Per dati sintetici, invece, consistono in informazioni generate artificialmente che riproducono le proprietà statistiche dei dati reali. Vengono utilizzati in vari contesti, come l'addestramento di modelli di machine learning al fine di tutelare i dati personali, generandone di sintetici. I dati sintetici permettono di superare limitazioni legate alla disponibilità o alla "sensibilità" dei dati reali, offrendo un'alternativa per analisi e sviluppo senza compromettere la riservatezza delle informazioni personali.

mentre ne risultano esclusi quelli con funzioni di semplice editing o che non alterano significativamente i dati forniti dall'utente come input. Inoltre, chi utilizza sistemi IA per creare contenuti che somigliano a persone, oggetti o eventi reali (es. deep fake), ad esempio, deve indicare chiaramente che il contenuto è generato o manipolato artificialmente, tramite etichette che rivelano in modo inequivocabile l'origine artificiale del contenuto, proteggendo così i consumatori da possibili inganni.

L'AI Act, inoltre, prevede e caldeggia il coinvolgimento delle associazioni a tutela dei consumatori in alcune circostanze.

Innanzitutto, il considerando 121, fa riferimento alle norme armonizzate definite nel Regolamento (UE) 1025/2012, IA e ruolo delle associazioni a tutela dei consumatori, come strumento chiave per garantire la conformità tecnica alle disposizioni di un regolamento specifico e, al contempo, per promuovere innovazione, competitività e crescita nel mercato unico europeo. Le norme armonizzate, come definite nell'articolo2, punto 1, lettera c) del Regolamento (UE) n. 1025/2012, sono le "specifiche tecniche" (o standard) adottate da organismi di normazione europei su mandato della Commissione Europea. Queste norme sono considerate uno strumento per dimostrare la conformità ai requisiti essenziali stabiliti nei regolamenti europei, in quanto rappresentano lo "stato dell'arte" (cioè il livello tecnologico e organizzativo più avanzato disponibile). Per i fornitori (compresi quelli di intelligenza artificiale), l'adozione di norme armonizzate fornisce un mezzo semplificato per dimostrare che i loro prodotti o servizi rispettano i requisiti del regolamento in questione. Questa conformità presunta riduce i costi amministrativi e offre certezza giuridica nel mercato unico, facilitando la libera circolazione dei prodotti e servizi. L'AI Act evidenzia, al considerando 121, la necessità di un processo inclusivo nell'elaborazione delle norme armonizzate che preveda il coinvolgimento di tutti i portatori di interessi pertinenti nell'elaborazione delle norme, in particolare, oltre alle organizzazioni dei consumatori, anche le PMI e i portatori di interessi in materia sociale e ambientale.

Il successivo considerando 142, inoltre, prevede un ulteriore momento di coinvolgimento di esperti in materia di tutela dei consumatori. Al fine di garantire che l'IA porti a risultati vantaggiosi sul piano sociale e ambientale, si promuove la ricerca e lo sviluppo di soluzioni di intelligenza artificiale che siano vantaggiosi dal punto di vista sociale ed ambientale, come le soluzioni basate sull'IA per aumentare l'accessibilità per le persone con disabilità o per affrontare le disuguaglianze socioeconomiche. È proprio nell'ambito di tali progetti che l'AI Act prevede che ci si debba basare sul principio della cooperazione interdisciplinare tra sviluppatori dell'IA, esperti in materia di disuguaglianza e non discriminazione, accessibilità e diritti ambientali, digitali e dei consumatori, nonché personalità accademiche.

Anche il considerando 165 prevede il coinvolgimento delle organizzazioni per la tutela dei consumatori al fine di garantire l'adozione nell'Unione di un'IA etica e affidabile. A tal riguardo si prevede che i fornitori di tutti i sistemi e modelli di intelligenza artificiale dovrebbero essere incoraggiati ad applicare, oltre che ai codici di condotta specifici sulla IA etica e affidabile, anche un insieme di misure ulteriori, tra le quali il coinvolgimento delle organizzazioni per la tutela dei consumatori nella progettazione e nello sviluppo dei sistemi di IA.

9. AI impiegate contro i diritti o interessi dei consumatori

Mentre, da un lato, l'intelligenza artificiale offre innumerevoli benefici e vantaggi – che approfondiremo nel capitolo seguente – che vanno dall'automazione di processi all'ottimizzazione dei servizi e alla personalizzazione della *user experience*, essa presenta, per contro, anche rischi significativi. Uno degli aspetti che preoccupa maggiormente è l'uso dell'IA contro i diritti e gli interessi dei consumatori, nonché l'incremento di crimini informatici che colpiscono le persone fisiche. I consumatori si

trovano ad affrontare una serie di minacce che vanno dalla manipolazione dei dati personali all'esposizione a frodi sofisticate e, spesso, i criminali informatici sfruttano proprio l'intelligenza artificiale per potenziare le loro attività illegali, rendendo gli attacchi maggiormente efficaci e difficili da rilevare. Al contempo, alcune aziende utilizzano l'IA in modi che possono ledere i diritti dei consumatori, ad esempio attraverso pratiche discriminatorie, invasione della privacy o manipolazione comportamentale.

In questo capitolo esamineremo alcune degli scenari in cui l'IA venga impiegata controi diritti e gli interessi dei consumatori e analizzeremo i crimini informatici maggiormente rivolti contro le persone fisiche. Attraverso riferimenti concreti e l'elenco dei temi da sviluppare, si intende fornire una panoramica esaustiva delle sfide attuali e, nel capitolo successivo, delle possibili soluzioni.

9.1 Panoramica sull'IA e i consumatori

Dagli assistenti virtuali come Siri e Alexa, alle raccomandazioni personalizzate su piattaforme come Netflix e Amazon, l'IA è diventata un componente essenziale dei prodotti e servizi moderni. Le applicazioni dell'IA nei settori dell'e-commerce, dei servizi finanziari, della sanità e dei trasporti hanno trasformato il modo in cui i consumatori interagiscono con il mondo digitale e fisico. Nel settore dell'e-commerce, l'IA viene utilizzata per analizzare il comportamento degli utenti, prevedere le preferenze dei consumatori e personalizzare l'esperienza di acquisto. Gli algoritmi di raccomandazione suggeriscono prodotti basati sulla cronologia di navigazione e di acquisto, aumentando le probabilità di vendita e migliorando la soddisfazione del cliente. Le banche e le istituzioni finanziarie utilizzano l'IA per rilevare frodi, valutare il rischio creditizio e offrire servizi personalizzati. Chatbot e assistenti virtuali forniscono supporto ai clienti 24 ore su 24, 7 giorni su 7, migliorando l'accessibilità dei servizi finanziari. Nella sanità, l'IA aiuta nella diagnosi precoce di malattie, nell'analisi delle immagini mediche e nella personalizzazione dei trattamenti. Le applicazioni di IA

possono analizzare grandi quantità di dati clinici per identificare pattern che potrebbero sfuggire all'occhio umano. Il settore dei trasporti ha visto l'introduzione di veicoli autonomi e sistemi di navigazione avanzati che utilizzano l'IA per migliorare la sicurezza stradale e ottimizzare i percorsi. Servizi come Uber e Lyft utilizzano algoritmi di IA per abbinare conducente e passeggero in modo efficiente.

L'adozione diffusa dell'IA ha anche avuto un impatto significativo sulla società e sulla cultura. Ha cambiato il modo in cui comunichiamo, consumiamo media e gestiamo le nostre finanze. Le piattaforme di social media utilizzano l'IA per curare i feed[16] di notizie, influenzando le informazioni a cui siamo esposti. Questo ha implicazioni non solo per le preferenze individuali, ma anche per il discorso pubblico e la formazione delle opinioni.

Con l'accessibilità crescente di strumenti e piattaforme basati sull'IA, anche le piccole imprese e i singoli consumatori hanno la possibilità di sfruttare queste tecnologie. Ad esempio, applicazioni per la traduzione linguistica, il riconoscimento vocale e l'editing fotografico avanzato sono ora disponibili a portata di mano attraverso smartphone e dispositivi personali.

L'IA, quindi, consente di creare esperienze altamente personalizzate, adattando prodotti e servizi alle esigenze specifiche di ogni consumatore. Questo aumenta la soddisfazione del cliente e può portare a scelte più informate.

Accanto ai benefici possiamo trovare anche i rischi connessi a tale nuova tecnologia. La raccolta e l'analisi di grandi quantità di dati personali sollevano seri problemi di

[16] Un feed di notizie è un flusso continuo di aggiornamenti, articoli o contenuti provenienti da varie fonti online, presentati in un formato organizzato e facilmente accessibile. Solitamente, i feed di notizie sono generati automaticamente da algoritmi che raccolgono e ordinano informazioni basandosi su criteri come interessi personali, cronologia di navigazione o rilevanza rispetto al profilo dell'utente. I feed di notizie si trovano in vari contesti, come social network, siti di notizie o "feed RSS" (un formato specifico per ricevere aggiornamenti da siti web o blog). L'obiettivo di un feed di notizie è permettere agli utenti di accedere rapidamente ai contenuti aggiornati senza doverli cercare manualmente.

tutela della riservatezza e dei dati personali. I consumatori sono spesso inconsapevoli della portata delle informazioni raccolte su di loro e di come queste vengono utilizzate.

Gli algoritmi di IA, inoltre, possono perpetuare o amplificare bias[17] esistenti nei dati di addestramento, portando a decisioni discriminatorie nei confronti di certi gruppi. Si pensi, ad esempio, a quei sistemi di valutazione del credito, gestiti da una IA, che penalizzano – in modo non voluto né previsto dagli utilizzatori della IA – individui provenienti da determinate aree geografiche o con specifici profili socioeconomici.

Certi sistemi di IA, ancora, abbiamo visto che hanno il loro "motore" in algoritmi talmente complessi che risulta spesso impossibile comprendere quale percorso "decisionale" abbia seguito la IA per giungere a un determinato risultato. Parliamo, in questi casi, di "black box", scatole nere dai meccanismi insondabili. Ebbene, in questi casi è difficile (se non impossibile), per i consumatori, comprendere come vengono prese le decisioni che li riguardano.

Un ulteriore aspetto potenzialmente negativo lo potremmo ricondurre ad un'eccessiva fiducia che il consumatore dovesse sviluppare nei confronti dell'IA: tale dipendenza tecnologica sarebbe in grado di ridurre la capacità dei consumatori di prendere decisioni autonome, critiche o obiettive nei confronti di prodotti o servizi, o di svolgere compiti senza assistenza.

La progressiva compenetrazione delle tecnologie guidate da IA, nei dispositivi domestici e personali aumenta la superficie di attacco per i cybercriminali. In altre parole: più sono i nostri dispositivi connessi alla nostra rete o ai nostri account, maggiori sono le opportunità, per i criminali informatici, di individuare una vulnerabilità da sfruttare per compromettere la nostra vita "digitale". Si pensi, ad

[17] Bias (termine anglosassone che significa "pregiudizio") nell'ambito dell'intelligenza artificiale si riferisce a pregiudizi, distorsioni o errori sistematici che influenzano i risultati prodotti da un sistema o un modello di AI. Questi bias possono emergere in diverse fasi dello sviluppo e dell'uso di un sistema di AI, come nella raccolta dei dati, nella progettazione dell'algoritmo o nell'interpretazione dei risultati.

esempio, ai dispositivi IoT (*Internet of Things*) non adeguatamente protetti (tra i quali possiamo far rientrare le smart TV, gli assistenti virtuali, le telecamere smart, i sistemi di videocitofono intelligenti, gli smartwatch, gli elettrodomestici da cucina wireless, l'illuminazione intelligente, le automobili intelligenti etc.) possono essere sfruttati dagli attaccanti malevoli, compromettendo la sicurezza domestica o la privacy degli utenti.

L'IA, pertanto, offre opportunità senza precedenti per migliorare la vita dei consumatori, ma richiede anche una gestione attenta dei rischi associati. È fondamentale che le aziende, i legislatori e i consumatori stessi lavorino insieme per massimizzare i benefici e minimizzare i potenziali danni. I consumatori, in particolare, devono essere informati e consapevoli dei rischi e delle opportunità legate all'IA che decidono di impiegare. Ciò significa curare l'alfabetizzazione digitale personale, su come funziona IA, su come vengono utilizzati i propri dati e su come convivere con queste tecnologie minimizzando i danni potenziali. Inoltre, l'alfabetizzazione consentirà loro di effettuare scelte consapevoli, previa valutazione delle implicazioni delle tecnologie che si utilizzano. Infine, nel caso in cui dovessero ravvisare abusi dovrebbero affidarsi alle organizzazioni a tutela dei consumatori, segnalando le circostanze, in modo da trasformare quella che può essere un'esperienza personale negativa in un supporto alla prevenzione per tutti i consumatori.

9.2 Gli impieghi malevoli della IA nei confronti dei consumatori

a) Tecniche d'attacco come mattoncini compatibili

Le tecniche di attacco informatico hanno subito un'ulteriore evoluzione con l'avvento dell'IA. Le singole tecniche di attacco possono essere paragonate a dei mattoncini che gli attaccanti modulano e assemblano in modi diversificati, più o meno sofisticati, per

orchestrare attacchi mirati ai consumatori. A ciascun mattoncino corrisponde un elemento specifico dell'attacco nella sua complessità—che sia una vulnerabilità del sistema, un exploit software, o una strategia di ingegneria sociale—e la loro combinazione permette di creare l'architettura dello specifico attacco, in modi sempre più personalizzati e difficili da prevedere. Il fatto che questi elementi si prestino ad essere modulati tra loro, ci consentono di descrivere il singolo attacco come una concatenazione, più o meno articolata, di specifiche tecniche di attacco. Le tipologie di attacco diventano quindi estremamente variabili, poiché la disposizione dei mattoncini (o anelli della catena) può essere modificata per aggirare nuove misure di sicurezza o per sfruttare debolezze specifiche di un obiettivo. Ad esempio, un attacco può iniziare con una semplice e-mail di phishing, magari potenziata dall'IA, per analizzare le risposte dell'utente e, successivamente, utilizzare malware avanzati per ottenere accesso non autorizzato ai dati sensibili. In altri casi, invece, le prime tecniche di attacco sono rappresentate da pratiche finalizzate a ottenere informazioni sulla vittima, che successivamente verranno impiegate per aggredire o violare un determinato sistema informatico, attraverso le ulteriori tecniche di ingegneria sociale (di cui si tratterà di seguito). L'evoluzione tecnologica gioca un ruolo cruciale in questo panorama in continua mutazione. Man mano che emergono nuove tecnologie—come l'Internet delle cose (IoT), il cloud computing e le reti 5G—si aprono anche nuove superfici di attacco. Gli aggressori spesso sfruttano anche le novità nel campo della tecnologia per sviluppare nuovi metodi di attacco. L'IA stessa, mentre da un lato offre strumenti efficaci per la difesa cibernetica, dall'altro viene utilizzata per automatizzare attacchi, creare malware adattivi o condurre campagne di disinformazione su larga scala. In tale contesto, la sicurezza informatica non può essere considerata un obiettivo statico, ma piuttosto un processo dinamico che richiede un adattamento costante: un moderno gioco di "guardie e ladri" sul versante tecnologico. Le

organizzazioni e gli individui devono essere consapevoli che le minacce odierne non solo sono più numerose rispetto al passato, ma anche più sofisticate e personalizzate.

Le **modalità** attraverso le quali una tipologia di attacco informatico può manifestarsi dipende da una molteplicità di fattori che influenzano sia la scelta delle tecniche utilizzate sia l'efficacia dell'attacco stesso. Le principali variabili che entrano in gioco possono dipendere dalla tecnologia disponibile, dalle informazioni a disposizione dell'attaccante sulla vittima, dal contesto personale o socio-economico e così via. Qualora, ad esempio, la vittima dovesse disporre di un gran numero di dispositivi eterogenei connessi alla propria rete locale, l'attaccante avrebbe a disposizione una superficie d'attacco più ampia anche in considerazione del fatto che maggiori sarebbero le possibilità di individuare una qualche vulnerabilità in uno dei tanti dispositivi connessi alla rete della vittima. Qualora, ancora, l'attaccante dovesse disporre di un numero significativo di informazioni sulla vittima, potrebbe impiegare le stesse in un attacco di spear phishing. Nei casi in cui le informazioni sulla vittima fossero assolutamente necessarie per passare alle ulteriori fasi dell'attacco, allora l'attaccante impiegherà tecniche espressamente orientate all'ottenimento di tali informazioni (con tecniche di ingegneria sociale o infettando i dispositivi della vittima con malware del tipo "infostealer"[18]).

[18] Un *infostealer* è un tipo di malware progettato per raccogliere informazioni sensibili da un dispositivo infettato. Questo software malevolo è programmato per rubare dati personali, come credenziali di accesso (username e password) da browser, client email o applicazioni; dati bancari o finanziari; altre informazioni memorizzate nei browser (come cookie, dati delle carte di credito o cronologia); file o documenti archiviati sul dispositivo; token di accesso a servizi (come social media, vpn o piattaforme di lavoro); screenshot o registrazioni dell'attività sullo schermo. L'infostealer viene diffuso attraverso tecniche come phishing, download dannosi, allegati email o vulnerabilità di software non aggiornati e, una volta installato, analizza il dispositivo per identificare e raccogliere le informazioni. In questo caso, notiamo, che solo nell'attacco finalizzato alla raccolta delle informazioni possono essere impiegate due tattiche d'attacco: il mattoncino iniziale del phishing per veicolare il malware (infostealer, in questo caso) e il mattoncino successivo, rappresentato dall'infostealer, per acquisire le informazioni. Queste ultime potranno rappresentare un ulteriore "mattoncino" da impiegare per ulteriori fasi d'attacco, quali, ad esempio, quelle di ingegneria sociale (potendo essere utili, ad esempio, a infondere nella vittima una sensazione di "serenità". Si pensi, ad esempio, a un soggetto che richieda informazioni ad una vittima, dimostrando la propria legittimazione ad ottenere quelle informazioni proprio sulla base delle informazioni precedentemente acquisite. Per semplificare ulteriormente: la vittima alla quale viene richiesto di fornire ulteriori dati sensibili potrebbe essere indotta a

Anche le **motivazioni** dietro l'attacco possono determinare le modalità d'attacco e le tecniche adottate dall'attaccante: se l'obiettivo è il guadagno economico, l'attaccante potrebbe utilizzare ransomware (al fine di estorcere del denaro) o rubare le informazioni finanziarie della vittima (al fine di accedere direttamente ai conti della vittima) o, ancora, usare le classiche frodi finanziarie, anche tramite l'illusione di facili guadagni (un po' dei moderni Gatto e Volpe che convincono Pinocchio che convenga investire nell'albero degli zecchini. Se, invece, intendesse danneggiare la reputazione della vittima, potrebbe optare per la diffusione di informazioni false o sensibili.

Tra le ulteriori variabili in grado di incidere sulle tecniche impiegate per l'attacco vi sono anche le **competenze** di cui disponga l'attaccante (un attaccante con risorse, competenze ed esperienza può orchestrare attacchi più elaborati e difficili da rilevare), la "vulnerabilità" del consumatore. Queste ultime possono riguardare sia i sistemi informatici della vittima, sia la vittima stessa. Nel caso in cui ad essere deboli siano i sistemi informatici, dal punto di vista della sicurezza intrinseca (pensiamo a un sistema mal configurato che lascia aperte delle porte d'accesso ai malintenzionati, oppure a un sistema informatico non adeguatamente protetto da antivirus o software anti-intrusione, un sistema operativo non aggiornato etc.) gli attaccanti potrebbero sfruttare queste debolezze a loro vantaggio. Allo stesso modo, qualora la debolezza riguardasse la vittima – perché, ad esempio, manca di consapevolezza circa le potenziali minacce che affliggono il proprio sistema informatico o, ancora, perché si trova in una situazione di vulnerabilità determinata da una crisi economica che lo

fidarsi, poiché l'attaccante dimostra di conoscere già alcune informazioni personali, come dettagli della sua attività lavorativa, password parzialmente mascherate o altre informazioni specifiche. Questo meccanismo sfrutta il senso di sicurezza percepito dalla vittima, spingendola a collaborare con l'attaccante, magari credendolo un tecnico, un collega o un'autorità legittimata.

Una volta esfiltrati i dati tramite l'infostealer, questi vengono inviati a un server remoto controllato dagli attaccanti, spesso tramite connessioni cifrate per evitare il rilevamento. Tali dati potranno essere reimpiegati in diversi modi: potranno essere venduti sul dark web o utilizzati direttamente per frodi, accessi non autorizzati o altri attacchi mirati. Tra gli infostealer più conosciuti si segnalano RedLine Stealer (usato dai cybercriminali per rubare dati di login e dettagli di pagamento), Agent Tesla (specializzato nel furto delle informazioni da applicazioni e client di posta elettronica) o, ancora, Raccoon Stealer (impiegato nel furto di credenziali, dati bancari e altre informazioni memorizzate nel sistema).

induce a trovare una rapida ed urgente soluzione ai propri problemi economici – l'attaccante potrebbe, ancora una volta, sfruttarla a proprio vantaggio. Una vittima che si trovi in uno stato di difficoltà economica, sia pur momentanea, sarebbe maggiormente esposta, infatti, a truffe legate a offerte di lavoro o aiuti finanziari.

Anche il **tempo** è un fattore determinante nella riuscita dell'attacco e, conseguentemente, nel danno che una vittima potrebbe patire in conseguenza dell'attacco. Qualora, infatti, l'attaccante sia disposto a investire più tempo nell'attacco, potrebbe concentrarsi maggiormente sulla pianificazione dell'attacco, rendendo quest'ultimo maggiormente efficace e, potenzialmente, più pericoloso per la vittima. Si pensi, ad esempio, all'attaccante che "investe" molto del tempo dell'attacco (e spesso si parla di settimane o mesi di preparazione dell'attacco) nella fase di raccolta delle informazioni sulla vittima, ad esempio installando dei software-spia (spyware, RAT, trojan) all'interno dei sistemi della vittima e stando, silenzioso, ad osservare le operazioni compiute dalla vittima attraverso il proprio sistema informatico.

b) Gli attacchi arricchiti dall'Intelligenza Artificiale

La criminalità informatica trova nell'intelligenza artificiale un valido supporto per i propri attacchi. Abbiamo già visto, in precedenza, come uno strumento di per sé neutro, possa essere impiegato per le più disparate finalità. I sistemi di intelligenza artificiale maggiormente noti al grande pubblico sono, naturalmente, progettati per impedire all'utente comune di ottenere informazioni malevole o per impedirgli, comunque, di ottenere risposte finalizzate a portare a segno un qualche attacco informatico. Tuttavia, potrebbero individuarsi delle modalità per indurre la IA a rilasciare, comunque, tale tipo di informazioni (tramite tecniche di *jailbreaking*[19]) o,

[19] Il termine jailbreaking (letteralmente "evasione dalla prigione" o "rottura della prigione") è un neologismo composto da due elementi: "jail" (prigione o carcere) e "breaking" (dal verbo "to break", rompere o sfuggire). Originariamente, il termine è stato usato per indicare il processo di rimozione delle restrizioni imposte da un produttore su un dispositivo elettronico (come iPhone, iPad, console di gioco, ecc.),

ancora, impiegando sistemi di intelligenza artificiale sui quali gli sviluppatori non abbiano imposto delle limitazioni specifiche (quali quelle, ad esempio, finalizzate a non realizzare immagini pornografiche o a rilasciare informazioni su tecniche di attacco informatico che potrebbero essere sfruttate anche dall'utente comune). In questo caso, impiegando cioè sistemi di intelligenza artificiale privi di limitazioni in tal senso imposte dagli sviluppatori, i sistemi di intelligenza artificiale possono rappresentare un pericoloso alleato dei cybercriminali.

L'intelligenza artificiale può essere impiegata dai criminali informatici per diverse finalità e in ragione di alcune potenzialità che solo uno strumento come l'intelligenza artificiale può offrire loro. La IA consente, innanzitutto, una **maggiore velocità nell'esecuzione degli attacchi**, poiché permette di automatizzare processi che prima richiedevano tempo e risorse umane significative. Questo porta a un **aumento della scalabilità**[20], consentendo ai criminali informatici di colpire un numero molto più elevato di vittime simultaneamente.

Un'altra caratteristica fondamentale degli attacchi "aumentati" dalla IA è la **sofisticazione**: l'IA permette di sviluppare malware e tecniche di intrusione più avanzate, capaci di eludere i sistemi di sicurezza tradizionali. Anche i sistemi di autenticazione o di sicurezza basati sulla biometria (si pensi, ad esempio, al

permettendo agli utenti di ottenere un controllo completo sul sistema operativo e di installare software non autorizzato o modificare impostazioni protette. In senso più ampio, "jailbreaking" può essere usato metaforicamente per riferirsi all'azione di superare o aggirare restrizioni in un sistema qualsiasi, incluso il contesto delle intelligenze artificiali. Il jailbreaking dei sistemi di intelligenza artificiale si riferisce a tecniche utilizzate per aggirare le restrizioni o i vincoli di sicurezza implementati in modelli di intelligenza artificiale, come chatbot o assistenti virtuali. Questi sistemi sono progettati per operare entro limiti ben definiti, come evitare contenuti offensivi, non violare disposizioni legali o etiche e rispettare la privacy degli utenti. Tuttavia, il jailbreaking mira a forzare il sistema a violare queste restrizioni, spesso sfruttando vulnerabilità nel suo design. l processo di jailbreaking può avvenire attraverso diversi metodi, tra cui il prompt engineering manipolativo, in cui l'utente crea input particolarmente elaborati che confondono il modello, inducendolo a generare risposte altrimenti proibite.

[20] La scalabilità è un concetto fondamentale nell'informatica ed è usata per descrivere la capacità di un sistema, rete, processo o organizzazione di crescere ed adattarsi alle esigenze crescenti o variabili senza compromettere significativamente le prestazioni, la funzionalità o la qualità.

riconoscimento facciale, tramite impronte digitali o riconoscimento della voce) possono essere elusi con strumenti generati dall'intelligenza artificiale.

Gli attacchi diventano inoltre più **personalizzati**, in ragione dell'analisi delle informazioni relative alle vittime. Ciò aumenta l'efficacia di pratiche come il "phishing mirato" o *spear phishing*[21].

Anche il "tradizionale" malware può divenire maggiormente insidioso con l'intelligenza artificiale. Il tradizionale attacco tramite ransomware, infatti, funziona crittografando i dati di una vittima con il ransomware stesso e richiedendo un riscatto per la loro decifratura. Tuttavia, l'integrazione dell'intelligenza artificiale porta questo tipo di attacco a un livello completamente nuovo, migliorandone l'efficienza e l'impatto. Grazie all'IA, ad esempio, un ransomware "intelligente" può analizzare rapidamente i sistemi della vittima e identificare i file più critici e preziosi (quali documenti finanziari, database aziendali, progetti riservati, backup strategici e così via). Il ransomware, quindi, può ignorare file inutili o di bassa priorità, concentrandosi esclusivamente sui dati il cui mancato accesso potrebbe avere un impatto devastante sulla vittima: questo approccio ottimizza l'efficacia dell'attacco e aumenta le probabilità che la vittima consideri il pagamento come l'unica soluzione. Inoltre, un ransomware intelligente potrebbe impiegare le tecniche di analisi predittiva per determinare l'importo ideale da richiedere come riscatto. Dopo aver analizzato fattori come la situazione finanziaria della vittima (tramite informazioni pubbliche o rubate), dimensione e il settore aziendale, la valutazione del "possibile" valore dei dati crittografati, il malware suggerisce di chiedere, come riscatto, un importo che sia sufficientemente alto da essere redditizio per l'attaccante, ma non così alto da scoraggiare il pagamento da parte della vittima.

[21] Lo spear phishing è una forma mirata di phishing, in cui un attaccante invia messaggi fraudolenti altamente personalizzati a una persona o a un gruppo specifico con l'obiettivo di indurre la vittima a compiere un'azione dannosa, come cliccare su un link malevolo, fornire informazioni sensibili o scaricare malware.

L'intelligenza artificiale può aiutare, inoltre, i criminali informatici ad eludere le misure di sicurezza predisposte dalle vittime. L'IA offre anche una **maggiore capacità di adattamento:** gli algoritmi possono modificare dinamicamente il comportamento degli attacchi in risposta alle difese incontrate, migliorando la probabilità di successo. Questa adattabilità si estende all'elusione dei sistemi di rilevamento, poiché i malware possono mutare e imparare dai tentativi falliti, rendendo più difficile la loro individuazione. Infatti, i malware potrebbero essere dotati della capacità di analizzare l'ambiente in cui si trovano e adattare il loro comportamento per eludere *sandbox*[22] e sistemi di rilevamento.

Una delle caratteristiche essenziali dell'IA, oltre alla rapidità di esecuzione delle operazioni, è quello di offrire la capacità di una **profonda analisi dei dati**, permettendoai cybercriminali di identificare vulnerabilità non ancora note o di sfruttare in modo più efficiente quelle esistenti. Le vulnerabilità "*zero-day*" possono essere individuate, infatti, da sistemi di intelligenza artificiale. L'eventuale scoperta di zero day dà all'attaccante un vantaggio formidabile nei confronti della vittima, posto che per tale vulnerabilità non sono state individuate ancora le contromisure (proprio perché la vulnerabilità è nota solo all'attaccante che la sfrutta, mentre è sconosciuta al resto del mondo).

Tutte le funzionalità offerte dall'IA, inoltre, sono spesso semplici da gestire tanto che si **riduce la necessità che il criminale informatico sia dotato di competenze avanzate.**

Infine, l'uso di tecniche come i *deepfake* **aumenta la verosimiglianza** delle frodi, facilitando l'inganno delle vittime attraverso contenuti multimediali falsificati ma estremamente realistici. Anche l'uso di algoritmi di elaborazione del linguaggio

[22] Una *sandbox* (che richiama alla mente i piccoli recinti riempiti con la sabbia al fine di consentire ai bambini di giocare in sicurezza, attutendone le eventuali cadute) è un ambiente isolato progettato per eseguire e testare applicazioni, codici o programmi in modo sicuro, senza rischiare di compromettere il sistema operativo principale o le risorse dell'utente. È un meccanismo fondamentale per garantire la sicurezza informatica e il controllo degli effetti di programmi potenzialmente dannosi o non fidati.

naturale (NLP) consente di generare testi di phishing in diversi linguaggi e con formule altamente personalizzate in modo da ottimizzare l'efficacia dello strumento. Oltre ai *deepfake*, ai quali abbiamo già fatto cenno, è possibile generare messaggi vocali impersonando persone di fiducia delle vittime. Gli stessi strumenti consentono di manipolare le informazioni o creare disinformazione, influenzando l'opinione pubblica o danneggiando la reputazione di individui e organizzazioni.

L'intelligenza artificiale, quindi, amplifica le capacità dei criminali informatici e spesso offre strumenti anche a chi, in assenza di tale strumento di potenziamento, non avrebbe avuto le **competenze** e i mezzi per rappresentare una minaccia effettiva nei confronti di consumatori e di organizzazioni.

Una semplificazione la si ottiene, inoltre, negli attacchi ai dispositivi IoT come smart TV, assistenti vocali, telecamere di sicurezza (sfruttandone le vulnerabilità per violare l'intera rete locale della vittima o, per creare delle *botnet*[23] o per accedere a risorse e informazioni sensibili). Ulteriore agevolazione che gli attaccanti possono trovare nella IA è quella del miglioramento delle capacità di **persistenza**[24]: questo consente agli attaccanti di introdurre dei malware nei sistemi informatici della vittima che restano inattivi per lunghi periodi, attivandosi solo quando le condizioni sono ottimali, aumentando così la probabilità di successo senza essere scoperti.

[23] Una botnet è una rete di computer o dispositivi connessi a Internet che sono stati compromessi da malware specifici per renderli controllabili da malintenzionati chiamati botmaster o bot herder. Ogni dispositivo infetto nella rete è definito bot (o zombie) e viene utilizzato per eseguire azioni senza il consenso del titolare.

[24] Nell'ambito della sicurezza informatica, la persistenza si riferisce alla capacità di un malware o di un attaccante di rimanere presente e operativo in un sistema compromesso per un lungo periodo, anche dopo riavvii, aggiornamenti o interventi di mitigazione.

9.3 Tipologie più comuni di crimini informatici e IA

È importante precisare che, sebbene possiamo individuare i reati informatici che colpiscono maggiormente i consumatori, non possiamo ignorare il fatto che i rischi informatici sono intrinsecamente **trasversali**. La persona che in alcuni momenti della giornata agisce come consumatore, in altri potrebbe ricoprire il ruolo di responsabile in un ufficio della pubblica amministrazione. Una minaccia informatica che lo colpisce nella sua sfera personale potrebbe estendere i propri effetti anche nell'ambito lavorativo, compromettendo così la sicurezza dell'ente presso il quale opera. Appare difficile, pertanto, connettere un rischio informatico soltanto a una categoria di soggetti. Di conseguenza, la distinzione tra rischi informatici per i consumatori e per le organizzazioni diventa sempre più sottile, sottolineando la necessità di un approccio integrato alla sicurezza informatica. Negli ultimi tempi, i trend della sicurezza informatica evidenziano un aumento di attacchi sofisticati rivolti ai consumatori. Phishing mirato (o spear phishing) è sempre più diffuso, utilizzando tecniche avanzate di ingegneria sociale e persino deepfake per ingannare le vittime e, quindi, come abbiamo già avuto modo di osservare, l'incidenza dei sistemi di intelligenza artificiale per finalità malevoli è in progressivo aumento. Il ransomware non colpisce più solo le grandi organizzazioni, ma anche gli utenti domestici, criptando i dati personali e richiedendo riscatti. Inoltre, c'è una crescita significativa di malware mobile e attacchi ai dispositivi IoT domestici, che spesso dispongono di misure di sicurezza inadeguate.

Cercheremo, quindi, di descrivere sia pur brevemente le tipologie di minacce informatiche che colpiscono maggiormente i consumatori. Tendenzialmente, i consumatori risultano appetibili per gli attori di un attacco informatico (normalmente su larga scala) per una serie di obiettivi che, seppur diversificati, tendono a convergere verso il profitto personale e l'accesso a informazioni sensibili.

Possiamo elencare gli scopi o **motivazioni** principali degli attaccanti che abbiano i consumatori come loro bersaglio:

1) **guadagno finanziario diretto** (truffe online, phishing e frodi con carte di credito, estorsioni post-ransomware);
2) **furto di dati personali** (queste informazioni possono avere un grande valore sul "mercato nero telematico" e, obiettivo finale degli attaccanti è, quindi, la commercializzazione – e quindi il guadagno finanziario indiretto – di queste informazioni o, ancora l'impiego di tali informazioni personali nella commissione di ulteriori reati);
3) **compromissione di dispositivi personali** (obiettivo è il reimpiego di tali dispositivi in botnet o per attacchi DDoS o, ancora, per il mining di criptovalute).

Le motivazioni che animano queste attività criminali sono, quindi, principalmente legate al profitto economico. La commissione di reati mediante la rete Internet o le tecnologie è diventata un'attività altamente remunerativa per la criminalità organizzata, con bassi rischi rispetto alle attività criminose tradizionali. La pretesa di anonimato nelle attività online e la possibilità di operare a livello globale rendono queste attività particolarmente attraenti per i malintenzionati.

Oltre al profitto, vi è anche l'intento di sfruttare i consumatori come punti di ingresso per attacchi più complessi rivolti ad aziende o enti governativi. Compromettendo l'account di un individuo che ha accesso a sistemi aziendali o istituzionali, gli attaccanti possono superare le difese perimetrali e accedere a informazioni sensibili su larga scala.

Infine, alcune attività possono essere motivate da ideologie politiche o sociali, come nel caso di attacchi hacktivisti. Tuttavia, queste motivazioni sono meno comuni rispetto all'obiettivo predominante del guadagno finanziario.

Comprendere gli obiettivi e le motivazioni dei criminali informatici è essenziale per sviluppare strategie efficaci di protezione. I consumatori devono essere consapevoli

dei rischi e adottare misure preventive, come quelle che sono illustrate in questa sia pur breve descrizione.

a) Truffe Online

Le truffe online includono una vasta gamma di attività fraudolente, come siti di e-commerce falsi, aste online fraudolente, offerte di lavoro fasulle o schemi di investimento che promettono rendimenti irrealistici, vendita di prodotti contraffatti o inesistenti e così via.

Indipendentemente dalla forma che assumono, le truffe online condividono alcuni elementi chiave. In primo luogo, si basano sull'**inganno** e sulla manipolazione psicologica. I truffatori sfruttano emozioni umane fondamentali come la paura, l'avidità, la curiosità o l'urgenza per indurre le vittime a compiere azioni impulsive, a cliccare su un link senza pensarci troppo. Spesso impersonano soggetti o entità affidabili, come istituti bancari, autorità governative o persone conosciute, per guadagnare credibilità e far abbassare lo scudo offerto dalla diffidenza. I criminali utilizzano messaggi persuasivi per richiedere informazioni personali, credenziali di accesso o trasferimenti di denaro. Spesso la richiesta appare innocua: “sto partecipando a un concorso a premi, puoi darmi il tuo voto?”. Ovviamente il link potrebbe essere usato per veicolare qualsiasi oggetto malevolo. Oppure: “per votare dovresti inviarmi lo screenshot del messaggio che riceverai al tuo telefono”. Magari, in questi casi, si tratta del codice del secondo fattore di autenticazione che consente al criminale di entrare nella disponibilità del nostro account social. Una volta ottenuto il controllo del nostro account social, lo utilizzerà per tentare di impersonare noi stessi al fine di truffare i nostri contatti social (i quali saranno più propensi ad acconsentire alle eventuali richieste degli attaccanti proprio perché penseranno di aver a che fare con noi). Inoltre, gli attaccanti fanno leva sulla mancanza di consapevolezza o sull'inesperienza tecnologica delle vittime, sfruttando tecniche avanzate di ingegneria sociale per superare le nostre difese naturali. Il complesso delle tecniche impiegate

per sfruttare le vulnerabilità umane è definito "*social engineering*" (ingegneria sociale), che sarà affrontato più avanti.

Le truffe online si presentano in molte forme, ma alcune tipologie sono particolarmente "famigerate" e meritano un'attenzione specifica.

Pensiamo, anzitutto, al phishing tradizionale, consistente nell'invio massivo di email o messaggi (anche su applicazioni di messaggistica istantanea come WhatsApp), apparentemente provenienti da fonti affidabili, a un numero indefinito di destinatari. I messaggi spesso contengono richieste urgenti di aggiornare informazioni personali o di risolvere problemi di sicurezza, indirizzando le vittime verso siti web falsi dove le informazioni inserite vengono rubate o dove vengono installati dei malware. Si è già fatto cenno anche allo spear phishing, variante del phishing, dove, però, i truffatori raccolgono informazioni specifiche sulla vittima per personalizzare i loro messaggi. Questo aumenta la credibilità della comunicazione e la probabilità che la vittima cada nell'inganno. Il phishing può essere veicolato anche tramite SMS, messaggerie istantanee o chiamate vocali o video. Tra le motivazioni che possono essere offerte alle vittime di phishing o spear phishing vi sono quelle di "supporto tecnico": i truffatori si spacciano per tecnici di aziende note, come Microsoft o Apple, e contattano le vittime affermando che il loro dispositivo ha un problema di sicurezza che, ovviamente, si offrono di correggere. L'obiettivo, in questi casi è ottenere l'accesso remoto al dispositivo o installare malware nel dispositivo della vittima. Un'altra delle motivazioni dei truffatori è l'apparente intenzione di instaurare relazioni affettive con le vittime, in quelle che sono definite le "truffe romantiche": attraverso siti di incontri o social media, i truffatori, mostrando interesse particolare nei confronti delle vittime, guadagnando la loro fiducia prima di richiedere denaro per emergenze personali o spese di viaggio. Un'altra truffa particolarmente diffusa è quella sulle prospettive di investimento o di acquisto di criptovalute. In questa tipologia di truffe si promettono alle vittime rendimenti finanziari elevati in breve tempo, spesso

attraverso investimenti in schemi piramidali o in criptovalute inesistenti. I truffatori spesso utilizzano testimonianze false di personaggi famosi (create tramite tecnologie deepfake) o documentazione manipolata e dati falsi per convincere le vittime ad "investire" del denaro che, invece, va a finire nelle mani dei truffatori. I deepfake sono usati anche in altre tipologie di truffe, oltre a quelle in cui personaggi pubblici promuovono investimenti fraudolenti. In altri casi, infatti, si usano videochiamate o chiamate artefatte, ad esempio di un dirigente, per autorizzare trasferimenti di denaro. Abbastanza comuni sono anche le truffe legate alla vendita online di beni o servizi, in cui i truffatori creano siti web o annunci falsi per prodotti a prezzi stracciati. Dopo il pagamento, il prodotto non viene mai consegnato o è di qualità inferiore rispetto a quanto promesso.

Esistono, ancora, le frodi tramite email aziendale. In questa tipologia di attacco, spesso definita "*mail in the middle*" (nota anche come "*CEO fraud*" o BEC "*business email compromise*"), gli attaccanti – dopo aver compromesso la casella email del dirigente di una società o dopo aver confezionato un indirizzo email fasullo ma facilmente confondibile con quello del dirigente o dopo aver, comunque, confezionato un'email o una telefonata o una videochiamata (tramite algoritmi di deepfake) – rivolgono a un dipendente dell'azienda la richiesta di disporre un pagamento in favore di taluno (ovviamente in modo che gli attaccanti possano entrare nella disponibilità del denaro inviato). Un'altra variante consiste nell'alterazione del codice IBAN di una fattura inviata o ricevuta dalla vittima in modo da dirottarne il pagamento verso soggetti non legittimati.

Le truffe tradizionali, quindi, sfruttano le medesime metodologie di sempre ma con una veste tecnologica. La principale arma di **difesa** che i consumatori possono adottare richiede una maggiore consapevolezza dei rischi e l'adozione di buone pratiche di sicurezza.

Innanzitutto, è fondamentale sviluppare un atteggiamento critico e **diffidente** verso le comunicazioni non sollecitate. Diffidare di email, messaggi o chiamate che richiedono azioni immediate o che offrono vantaggi troppo allettanti. Occorre sempre diffidare dalle "offerte troppo belle per essere vere". Verificare sempre l'autenticità del mittente controllando gli indirizzi email o i numeri di telefono, e confrontandoli con quelli ufficiali disponibili sui siti web delle aziende e, magari, provando a richiamare i numeri "verificati" al fine di chiedere conferma della comunicazione ricevuta. Occorre diffidare anche quando l'attaccante, nella email dimostri di conoscere informazioni che pensiamo di conoscere solo noi: come una password. Se anche il truffatore dovesse dimostrare di essere nella disponibilità di una nostra password (attuale o che abbiamo avuto in precedenza), non significa che abbia già avuto accesso agli oggetti protetti da tale password: è ben possibile che la nostra password sia stata oggetto di un leak dopo essere stata trafugata dai database di uno dei servizi online da noi utilizzati. Consideriamo, poi, che l'effetto psicologico di indebolire la nostra capacità critica per indurci a cliccare su un link o ad altre attività rischiose, lo si potrebbe raggiungere anche nell'ipotesi in cui il criminale dimostri di conoscere una nostra password che avevamo già provveduto alla sostituzione.

Occorre, inoltre, **evitare di cliccare** su link o scaricare allegati da fonti sconosciute: se un messaggio sembra sospetto, è preferibile accedere al servizio in questione digitando manualmente l'indirizzo nel browser o utilizzando l'app ufficiale e non cliccando sul link offerto dal potenziale attaccante. Ciò in quanto l'indirizzo Internet (o URL) che ci è stato inviato e sul quale ci si chiede di cliccare, potrebbe essere stato confezionato in modo tale da farlo apparire come l'indirizzo "ufficiale"[25].

[25] Si pensi, ad esempio, alle tecniche di *typosquatting* (in cui si fanno piccoli cambiamenti nell'URL rispetto a quello originale, come cambiare una lettera o aggiungere/rimuovere un carattere: gooogle.com invece di google.com) o di *omografo phishing*, dove si sfruttano le somiglianze visive tra caratteri di alfabeti diversi (es. caratteri latini e cirillici): apple.com (con caratteri cirillici) invece di apple.com.

Altro aspetto fondamentale per prevenire queste tipologie di truffa è quello di **proteggere le proprie informazioni personali**. Non condividere mai credenziali di accesso, numeri di carte di credito o altri dati analoghi tramite email o messaggi. Le istituzioni affidabili (come, ad esempio, la nostra banca) non ci chiederà queste informazioni attraverso tali canali. Occorre essere cauti anche sui social media. È sempre meglio limitare le informazioni personali condivise pubblicamente e gestire con attenzione le impostazioni sulla privacy. I truffatori possono utilizzare queste informazioni per personalizzare i loro attacchi.

È, poi, necessario curare l'**aggiornamento** dei nostri dispositivi e dei software: installare regolarmente gli aggiornamenti del sistema operativo e delle applicazioni, e utilizzare software antivirus affidabili per prevenire infezioni da malware aiuta a ridurre il rischio.

Utilizzare l'**autenticazione a due fattori** (2FA) quando disponibile. Questo aggiunge un ulteriore livello di sicurezza agli account online, rendendo più difficile per i truffatori accedervi anche se riescono a ottenere la password.

b) Estorsioni tramite ransomware

Nel panorama delle minacce informatiche, i ransomware rappresentano una delle forme più pericolose e diffuse di attacchi ai danni dei consumatori. Questi malware criptano i dati presenti sui dispositivi delle vittime, rendendoli inaccessibili, e richiedono un riscatto in denaro, solitamente in criptovalute, per fornire la chiave di decrittazione.

Alla base di ogni attacco ransomware vi è l'infezione del sistema bersaglio attraverso un software malevolo progettato per criptare i dati. I metodi di diffusione più comuni includono email di phishing contenenti allegati infetti, download da siti web compromessi e sfruttamento di vulnerabilità non corrette nei sistemi operativi o nelle applicazioni. Una volta infiltrato nel sistema, il ransomware inizia a cifrare i file, spesso

utilizzando algoritmi di cifratura avanzati che rendono impossibile il recupero dei dati senza la chiave corretta.

I criminali dietro questi attacchi mirano a ottenere proventi sfruttando la necessità delle vittime di accedere ai propri dati: la formula è, sostanzialmente, quella del reato di estorsione. La richiesta di riscatto viene solitamente presentata attraverso un messaggio sul dispositivo infetto, indicando le istruzioni per il pagamento, spesso in criptovalute come Bitcoin, per garantire l'anonimato (o quantomeno una maggiore difficoltà di identificazione) dei malintenzionati.

Esistono diverse varianti di ransomware, alcune delle quali adottano tecniche particolarmente sofisticate per massimizzare l'efficacia dell'attacco e la probabilità di ricevere il pagamento del riscatto. Una di queste varianti è chiamata “double extortion”: il criminale, oltre a cifrare i documenti tramite il malware, li preleva dal sistema bersaglio e richiede alla vittima un ulteriore riscatto per non diffondere onlinei contenuti esfiltrati (questa tecnica è chiamata “*double extortion*” proprio perché, una somma di denaro è richiesta in cambio della chiave necessaria per decifrare i file infetti dal ransomware, e un’ulteriore somma è richiesta per non diffondere online i dati esfiltrati). Se i documenti rubati coinvolgono la sfera intima della persona, come immagini o video a sfondo sentimentale o sessuale, la minaccia di divulgazione può avere un impatto emotivo e psicologico ancora maggiore. Questo tipo di estorsione sfrutta la paura del danno reputazionale e personale per forzare la vittima al pagamento.

Un'altra evoluzione nel panorama dei ransomware è rappresentata dal modello "Ransomware as a Service" (RaaS). In questo scenario, sviluppatori di ransomware offrono le loro piattaforme e competenze a terzi, spesso criminali con limitate conoscenze tecniche, in cambio di una percentuale dei riscatti ottenuti. Questo modello di business criminale ha abbassato le barriere all'ingresso per lanciare attacchi ransomware (non essendo richiesta sostanzialmente alcuna competenza,

posto che il novello criminale potrebbe limitarsi ad acquistare un servizio in grado di automatizzare in modo estremamente semplificato l'intero attacco ransomware, dall'invio iniziale delle email di phishing alla fase di acquisizione dei proventi illeciti dalle vittime), aumentando il numero di attacchi e la loro diffusione globale.

Con l'avvento dell'intelligenza artificiale e dei deepfake, alcuni attacchi ransomware possono essere combinati con tecniche avanzate per ingannare ulteriormente le vittime. Ad esempio, i criminali possono utilizzare deepfake vocali o video per impersonare dirigenti aziendali o persone di fiducia, inducendo gli utenti a eseguire azioni che facilitano l'infezione da ransomware o il pagamento del riscatto.

La protezione dai ransomware richiede un approccio proattivo e multilivello, combinando misure tecniche, comportamentali e di consapevolezza.

Innanzitutto, è fondamentale mantenere aggiornati tutti i dispositivi e i software utilizzati. Gli **aggiornamenti** spesso includono patch di sicurezza che correggono vulnerabilità sfruttabili dai ransomware per infettare il sistema. L'installazione di un software **antivirus** affidabile e il suo costante aggiornamento possono rilevare e bloccare diverse varianti di ransomware. La **formazione** e la **consapevolezza** sono altrettanto importanti: molti ransomware si diffondono attraverso email di phishing, e per questo motivo è essenziale essere cauti nell'aprire allegati o cliccare su link ricevuti da mittenti sconosciuti o sospetti. La cautela di "non accettare caramelle dagli sconosciuti" è qualcosa che ci è stato insegnato fin da piccoli. Continuiamo a seguire tale consiglio anche online. Diffidando da ciò che appare troppo bello per essere vero (caramelle dagli sconosciuti) potremmo ridurre notevolmente le probabilità di essere vittime di questi o altri attacchi da parte della criminalità informatica. Occorre, pertanto, verificare sempre l'autenticità delle comunicazioni e adottare un atteggiamento critico e diffidente nei confronti di richieste inaspettate.

È necessario, poi, effettuare regolarmente il **backup** dei nostri dati su supporti esterni (rispetto ai sistemi informatici di cui facciamo il backup) o servizi cloud sicuri è una misura fondamentale. È importante, infatti, assicurarsi che i backup siano isolati dal sistema principale per evitare che vengano anch'essi criptati durante l'attacco. In caso di infezione da ransomware, disporre di copie aggiornate dei file permette di ripristinare le informazioni.

Implementare soluzioni di sicurezza avanzate, come firewall e sistemi di rilevamento delle intrusioni, può offrire una protezione aggiuntiva, soprattutto in ambito aziendale. L'utilizzo di strumenti di **filtraggio del traffico email e web** può ridurre il rischio di esposizione a contenuti malevoli.

Infine, in caso di attacco ransomware, è importante **non** cedere immediatamente alla pressione di **pagare il riscatto**. Non vi è alcuna garanzia che i criminali forniscano la chiave di decrittazione dopo il pagamento. Si tratta, pur sempre di criminali! Non possiamo certo aspettarci che rispettino i termini di un accordo. È consigliabile, invece, rivolgersi alle autorità competenti e a professionisti della sicurezza informatica per valutare le opzioni disponibili.

c) Furto d'identità

Il furto dell'identità consiste in quell'attività attraverso la quale il criminale informatico dopo aver raccolto il maggior numero di informazioni personali su un individuo le impiega direttamente per sostituirsi alla vittima ed accedere ai suoi conti bancari o, indirettamente, per commettere truffe o altre attività illecite (come aprire conti bancari, ottenere carte di credito, richiedere prestiti o effettuare acquisti a nome della vittima) avvalendosi dell'identità della vittima.

Le conseguenze possono essere devastanti, influenzando la vita finanziaria, professionale e personale delle vittime. Alla base di ogni furto di identità vi è l'acquisizione di informazioni personali, come dati anagrafici, codici identificativi

(come codice fiscale o numero di passaporto), dettagli bancari e credenziali di accesso a servizi online. I truffatori, come abbiamo visto, utilizzano queste informazioni per impersonare la vittima, accedere ai suoi conti, effettuare acquisti, ottenere prestiti o commettere reati a suo nome.

Due sono le principali modalità attraverso le quali i criminali ottengono queste informazioni: acquisizione tramite ricerche online e tecniche di OSINT (Open Source Intelligence) e acquisizione diretta presso la vittima attraverso tecniche di ingegneria sociale.

Le tecniche di OSINT si basano sulla raccolta di informazioni da fonti pubblicamente accessibili. I criminali sfruttano la grande quantità di dati che sono quotidianamente condivise online per creare profili dettagliati delle loro vittime. Tra le fonti maggiormente impiegate per reperire informazioni sulle vittime troviamo:

- i social network (dettagli come data di nascita, luogo di residenza, luoghi frequentati, relazioni personali e professionali possono essere facilmente disponibili su tali piattaforme);
- blog e siti web personali (su tali piattaforme molte persone condividono storie personali, curriculum vitae, hobby e altri dettagli che possono essere utilizzati per ricostruire la loro identità);
- registri pubblici e documenti pubblicati dalle pubbliche amministrazioni (si pensi, ad esempio, alle informazioni pubblicate in “Amministrazione trasparente”, in base al D.Lgs. 33/2013 dalle pubbliche amministrazioni e da alcune società. In tali sezioni possono reperirsi dati personali spesso molto dettagliati come interi curriculum vitae di chi sia stato, ad esempio, consulente di qualche pubblica amministrazione);

- motori di ricerca (le ricerche con tecniche avanzate possono rivelare numerosi dettagli di una persona, soprattutto se questa ha una presenza online significativa).

L'ingegneria sociale, invece, coinvolge la manipolazione psicologica delle persone per ottenere informazioni o indurle a compiere azioni che compromettono la loro sicurezza.

La principale differenza tra le due modalità risiede nel coinvolgimento diretto della vittima: nelle ricerche di OSINT, il criminale raccoglie passivamente informazioni già disponibili pubblicamente e la vittima potrebbe non essere consapevole che i suoi dati sono stati raccolti e utilizzati in modo malevolo. Quando, invece, siano impiegate tecniche di ingegneria sociale, il criminale interagisce attivamente con la vittima, manipolandola per ottenere informazioni o indurla a compiere azioni dannose. La prevenzione richiede consapevolezza e scetticismo nelle interazioni.

I furti di identità, quindi, si manifestano in diverse forme, adattandosi continuamente alle nuove tecnologie e ai comportamenti dei consumatori.

Una delle forme più comuni è il furto di identità finanziaria. In questo caso, i criminali ottengono accesso alle informazioni bancarie o delle carte di credito della vittima. Possono effettuare transazioni non autorizzate, prelevare fondi o aprire nuovi conti a nome della vittima. Questo tipo di furto spesso avviene attraverso attacchi di phishing, dove la vittima viene indotta a inserire le proprie credenziali su siti web falsi che imitano quelli legittimi.

Il furto di identità digitale riguarda l'accesso non autorizzato agli account online della vittima, come email, social media o servizi di cloud storage. I criminali possono utilizzare queste piattaforme per diffondere malware, inviare messaggi fraudolenti ai contatti della vittima o accedere a ulteriori informazioni personali.

Proteggersi dai furti di identità richiede una combinazione di consapevolezza, pratiche sicure e vigilanza costante. Innanzitutto, è fondamentale proteggere le proprie informazioni personali ed evitare di condividere dati attraverso email o messaggi non sicuri e fornire tali informazioni solo su siti web affidabili e protetti. Quando si naviga online, assicurarsi che le connessioni siano sicure, identificabili dalla presenza del protocollo "https://" e dell'icona del lucchetto nella barra degli indirizzi.

Creare **password forti e uniche** per ciascun account è essenziale. Le password dovrebbero essere composte da una combinazione di lettere maiuscole e minuscole, numeri e simboli. L'uso di un gestore di password può aiutare a memorizzare in modo sicuro queste credenziali. Attivare l'autenticazione a due fattori aggiunge un ulteriore livello di sicurezza, richiedendo un secondo metodo di verifica oltre alla password.

Mantenere **aggiornati i dispositivi personali** è fondamentale. Installare regolarmente gli aggiornamenti del sistema operativo e delle applicazioni aiuta a correggere vulnerabilità che potrebbero essere sfruttate dai criminali. Utilizzare software antivirus e firewall affidabili offre una protezione aggiuntiva contro malware e tentativi di intrusione.

Essere **prudenti nelle comunicazioni** è altrettanto importante. Diffidare di email o messaggi che richiedono informazioni personali o finanziarie, anche se sembrano provenire da fonti affidabili. In caso di dubbi, contattare direttamente l'ente o l'azienda attraverso canali ufficiali per verificare la legittimità della richiesta.

Limitare le informazioni condivise sui social media può ridurre il rischio che i criminali raccolgano dati personali. Evitare di pubblicare dettagli come indirizzo di casa, numero di telefono, data di nascita o informazioni sulle proprie abitudini quotidiane. Configurare le impostazioni sulla privacy per controllare chi può vedere i propri post e le informazioni del profilo.

Monitorare regolarmente i propri conti bancari e le carte di credito permette di individuare tempestivamente transazioni sospette. Richiedere periodicamente il proprio rapporto di credito aiuta a verificare la presenza di conti o attività non autorizzate. Segnalare immediatamente qualsiasi anomalia alla banca o all'istituto finanziario competente.

Per quanto riguarda le minacce avanzate come i deepfake, è importante essere consapevoli della loro esistenza e potenziale uso fraudolento. Mantenere una **comunicazione diretta** con colleghi, amici e familiari può aiutare a verificare l'autenticità di richieste insolite. In ambito professionale, implementare procedure di verifica per autorizzazioni e transazioni finanziarie, come conferme vocali o firme digitali, può prevenire frodi basate su contenuti manipolati.

Infine, **educarsi e rimanere informati** sulle ultime tattiche utilizzate dai criminali informatici è fondamentale. Partecipare a programmi di formazione sulla sicurezza digitale e consultare fonti affidabili permette di aggiornare costantemente le proprie conoscenze e adottare misure preventive efficaci.

d) Attacchi di Ingegneria sociale

Il complesso delle tecniche utilizzate per sfruttare le vulnerabilità umane è definito **ingegneria sociale** (social engineering). Possiamo definire l'ingegneria sociale come l'arte di utilizzare l'interazione sociale per persuadere un individuo o un'organizzazione a soddisfare una specifica richiesta proveniente dall'attaccante, dove l'interazione, la persuasione o la richiesta coinvolgono un'entità collegata a un sistema informatico. Queste tecniche possono manipolare psicologicamente le vittime, inducendole a divulgare informazioni sensibili o a compiere azioni che compromettono la sicurezza dei loro dati.

Gli ingegneri sociali sfruttano la natura disponibile e fiduciosa che la maggior parte delle persone dimostra, oltre al fatto che molti non si aspettano di essere vittime di

attività manipolative e spesso agiscono in modo piuttosto incauto. In sostanza, si trattadi meccanismi comportamentali che possono essere impiegati per convincere oalterare la percezione del rischio o del pericolo da parte della vittima. Alcune tecnicheutilizzate dagli ingegneri sociali sono tipiche anche di altri settori, come quello pubblicitario o delle vendite.

Ad esempio, la **tecnica della reciprocità** prevede che un soggetto sia più incline a dare il proprio consenso a chi gli ha già offerto qualcosa (si pensi alle offerte di assaggi gratuiti che inducono chi ha provato il prodotto ad acquistarlo). Un'altra è la **tecnica della scarsità**, secondo la quale la percezione di rapido esaurimento di un bene aumenta il desiderio di procurarselo prima che diventi indisponibile (come nelle offerte "scade domenica" o "ultimi due posti disponibili"). Anche la pulsione a conformarsi alla maggioranza conferisce a un'offerta, presentata come "desiderata" da molti utenti, una maggiore appetibilità (basti pensare all'influenza che le recensioni positive possono avere sulle decisioni di acquisto online, anche se queste approvazioni o "like" sono spesso creati artificialmente).

Un'ulteriore tecnica è quella dell'**informazione placebo** (o tecnica della "mindlessness"): studi psicologici hanno dimostrato che le persone sono più propense ad acconsentire a una richiesta purché sia accompagnata da una motivazione, anche se tautologica.

Lo studio della vittima—delle sue pulsioni, paure, preoccupazioni, desideri o situazione familiare—permette all'attaccante di individuare e sfruttare la vulnerabilità umana per eseguire con maggiore successo l'attacco di ingegneria sociale. Una vittima in condizioni economiche difficili sarà più vulnerabile ad attacchi che le prospettano un guadagno consistente. Tra le vulnerabilità umane potenzialmente sfruttabili vi sono l'autorevolezza (una richiesta che sembra provenire da una figura di autorità ha maggiori probabilità di essere accettata), il senso di colpa, il panico, l'ignoranza, il desiderio sessuale, l'avidità e la compassione. Tipica in quest'ultimo caso è la

cosiddetta **truffa nigeriana**, in cui un presunto amico o conoscente invia un'email richiedendo denaro alla potenziale vittima, motivando la richiesta con il furto del portafogli. Oppure, ancora, il messaggio SMS del "figlio" che chiede aiuto perché ha smarrito lo smartphone, è in grave difficoltà e ha dovuto procurarsi un nuovo numero telefonico per contattare la potenziale vittima.

Alla luce di queste riflessioni, possiamo meglio comprendere quali differenti tecniche possano essere impiegate in un attacco di **phishing** o di **spear phishing**. Sebbene il phishing sfrutti alcune di queste tecniche, a differenza dello spear phishing non basa l'efficacia dell'attacco sulla conoscenza specifica della vittima, ma sui "grandi numeri". Nel phishing, infatti, si invia lo stesso messaggio a migliaia di potenziali vittime, con una percentuale di successo che dipende anche dalla qualità del messaggio. Lo spear phishing, invece, consiste in un attacco mirato a una vittima specifica sulla quale l'attaccante ha già acquisito una profonda conoscenza (tramite l'uso di tecniche specifiche per reperire informazioni personali), aumentando esponenzialmente la pericolosità e la probabilità di successo rispetto al "semplice" attacco di phishing.

Il ciclo di un attacco di ingegneria sociale può, in genere, essere suddiviso in quattro fasi:

1. **Fase 1 (ricerca delle informazioni)**. L'attaccante raccoglie dati sulla vittima, che possono essere reperiti in vari modi, anche online tramite tecniche di **OSINT** (Open Source Intelligence).
2. **Fase 2 (costruzione del rapporto di fiducia)**. L'attaccante stabilisce una relazione con la vittima, guadagnandone la fiducia attraverso interazioni studiate.
3. **Fase 3 (Sfruttamento della fiducia ricevuta)**. Una volta ottenuta la fiducia, l'attaccante manipola la vittima per ottenere informazioni sensibili o indurla a compiere azioni dannose.

4. **Fase 4 (Utilizzo delle informazioni e raggiungimento dell'obiettivo)**. L'attaccante utilizza le informazioni ottenute per raggiungere il suo scopo, che può essere di natura finanziaria, di accesso a sistemi o altro.

Le tecniche più comuni di ingegneria sociale hanno nomi strani e curiosi: **tailgating**, **piggybacking**, **baiting**, **pretexting**, **quid pro quo**, **diversion theft**, **dumpster diving** e **shoulder surfing**. Queste tecniche vengono utilizzate strategicamente nelle diverse fasi dell'attacco, adattandosi alle circostanze specifiche per massimizzare l'efficacia dell'inganno.

Per proteggersi dagli attacchi di ingegneria sociale, è fondamentale adottare una seriedi precauzioni:

- È necessario **informarsi** sulle diverse tecniche di ingegneria sociale e su come funzionano. Partecipare a programmi di formazione sulla sicurezza informatica può aumentare la capacità di identificare e resistere a tentativi di manipolazione.
- Bisogna **evitare di fidarsi ciecamente** di comunicazioni che richiedono informazioni personali o azioni immediate: verificare sempre l'identità del mittente attraverso canali ufficiali prima di fornire dati sensibili o di eseguire richieste insolite (se, ad esempio, ricevessimo un’email da un amico che ci chiede il compimento di una determinata azione, sarebbe opportuno chiamare il nostro amico al telefono per chiedere conferma di quanto ricevuto via email).
- Occorre **ridurre al minimo le informazioni personali condivise** sui social media e su altre piattaforme pubbliche. Questo rende più difficile per gli attaccanti raccogliere dati utili per personalizzare gli attacchi.
- **Diffidare** di email, messaggi o chiamate non sollecitate, soprattutto se contengono allegati o link sospetti. Evitare di cliccare su link o scaricare allegati da fonti non verificate.

- Mantenere **aggiornati** i dispositivi e il software per correggere eventuali vulnerabilità che potrebbero essere sfruttate dagli attaccanti.
- Abilitare **l'autenticazione a due fattori** sui propri account aggiunge un ulteriore livello di sicurezza, rendendo più difficile per gli attaccanti accedere ai dati anche se riescono a ottenere le credenziali. Oltre a ciò, bisogna assolutamente evitare di utilizzare la stessa password per diversi servizi (ad esempio le credenziali e le password usate per accedere al nostro profilo social non dovrebbero coincidere con le credenziali che usiamo in altri ambiti, come per registrarci al sito web del nostro market online o della nostra banca).
- Diffidiamo di offerte che sembrano troppo vantaggiose o urgenti (ricordiamo il consiglio “non accettare caramelle dagli sconosciuti”). Gli attaccanti spesso sfruttano l'avidità o il senso di urgenza per indurre le vittime a compiere azioni affrettate.

10. Per concludere

L'intelligenza artificiale rappresenta una delle più grandi rivoluzioni tecnologiche della nostra era, un universo di possibilità che si espande a ritmo crescente. È uno strumento potente, capace di trasformare radicalmente ogni aspetto della nostra società: dalla sanità all'educazione, dall'industria alla vita quotidiana. Tuttavia, come ogni strumento, l'IA non è intrinsecamente buona o cattiva. Tutto dipende dall'uso che se ne fa, dagli scopi che si perseguono e dalle mani in cui viene posta. È un'arma a doppio taglio che, se utilizzata con saggezza e consapevolezza, può portare enormi benefici, ma che, se impiegata irresponsabilmente, può creare rischi e danni significativi.

La trasparenza, l'etica e la regolamentazione sono le fondamenta per garantirne un uso responsabile, mentre l'educazione e l'alfabetizzazione digitale sono gli strumenti per mettere il controllo dell'IA nelle mani di tutti, non di pochi.

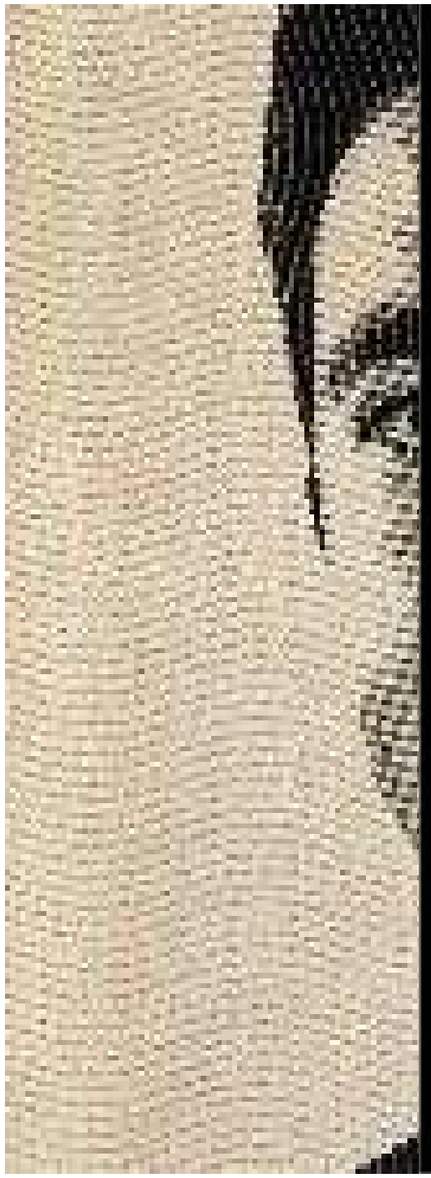
In fondo, la storia dell'umanità è sempre stata una storia di strumenti e di innovazioni che hanno ampliato i nostri orizzonti. L'intelligenza artificiale non fa eccezione. È un'opportunità straordinaria per dimostrare la nostra capacità di governare il cambiamento, di dirigere il progresso tecnologico verso un futuro che sia al servizio dell'uomo, dei suoi sogni e delle sue aspirazioni. Con studio, consapevolezza e un'etica solida, possiamo affrontare questa rivoluzione con speranza e determinazione, ricordandoci che, in definitiva, il destino è nelle nostre mani.

Francesco Paolo Micozzi

Avvocato cassazionista e Docente di Informatica Giuridica al Dipartimento di Giurisprudenza dell'Università degli Studi di Perugia. Componente dell'Academic Staff del Centro d'Eccellenza Jean Monnet "Building the Age of a Lawful and sustainable Data-Use" (BALDUS). Docente per la Scuola Superiore della Magistratura sui temi correlati alle indagini nell'ambito dei cybercrime.

Intelligenza artificiale. Usi distorti ai danni dei consumatori

Avv. Francesco Paolo Micozzi



Chi sono

- Avvocato cassazionista - A r r a y . e u
- Informatica Giuridica - UniPG
- Centro d’Eccellenza Jean Monnet - BALDUS - UniPG
- Module Leader - Jean Monnet - CIBER - UniPG
- Master I liv. “Cybercrime e Digital Forensics” - UniPG
- Consigliere dell’Ordine degli Avvocati di Cagliari
- CINI - Cyber Security National Lab - UniPG
- GdL della Fondazione Italiana per l'Innovazione Forense (FIIF) presso il CNF
- Commissione CNF per la valutazione degli avvocati specialisti in “Diritto dell'informazione, della comunicazione digitale e della protezione dei dati personali”

Alcune differenze

Furto (624 cp)	Estorsione (629 cp)	Truffa (640)
Chiunque s'impossessa della cosa mobile altrui, sottraendola a chi la detiene, al fine di trarne profitto per sé o per altri…	Chiunque, mediante violenza o minaccia, costringendo taluno a fare o ad omettere qualche cosa, procura a sé o ad altri un ingiusto profitto con altrui danno…	Chiunque, con artifizi o raggiri, inducendo taluno in errore, procura a sé o ad altri un ingiusto profitto con altrui danno…

Alcune tipologie di pericoli

Installazione/induzione all'installazione di software (infostealer, ransomware, rat…)

Richieste di invio di denaro

Richieste di registrazione a determinati servizi (ad esempio per poter partecipare alle elezioni politiche o amministrative)

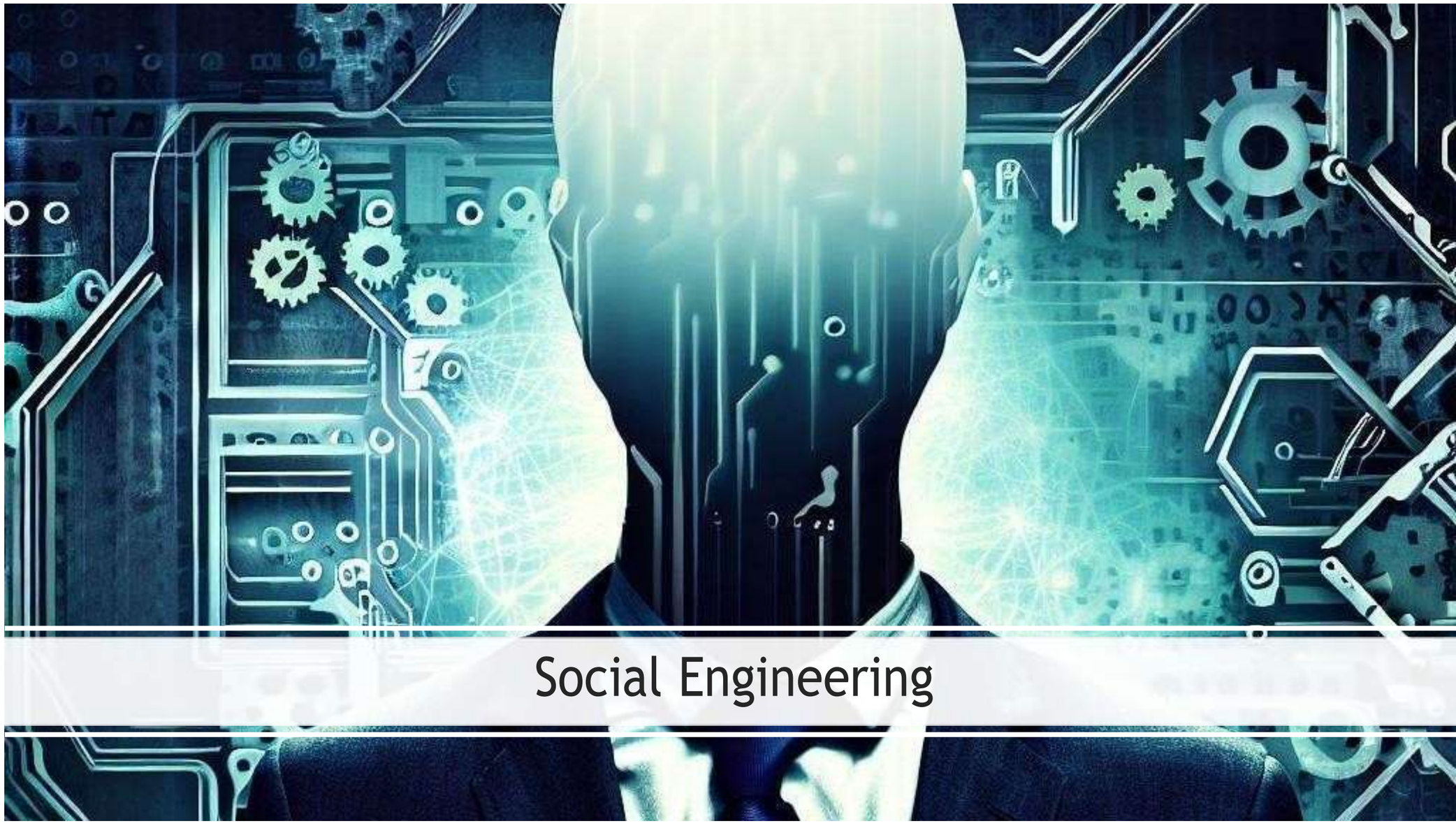
Richiesta di voti per la partecipazione a qualche concorso online

Richiesta di informazioni sulla propria identità personale (compresi documenti)

Richiesta di informazioni ricevute sul cellulare (ad esempio informazioni della 2fa)

Vendita di prodotti «per la salute o il benessere»

Vendita di prodotti «scontatissimi» su piattaforme «fake»



Social Engineering

Ingegneria Sociale:

“Ğhe science of using social inĞeracĞion as a means Ğo persuade an individual or an organisaĞion Ğo comply wiĞh a specific requesĞ from an aĞĞacker where eiĞher Ğhe social inĞeracĞion, Ğhe persuasion or Ğhe requesĞ involves a compuĞer-relaĞed enĞiĞy”

(Mouton et al., 2014)

Caratteristiche dell’ingegnere sociale

- Gli ingegneri sociali trasmettono sicurezza e controllo
- Gli ingegneri sociali offrono regali o favori gratuiti
- Gli ingegneri sociali usano l'umorismo
- Gli ingegneri sociali possono sempre fornire una motivazione

Hardware, software e... wetware

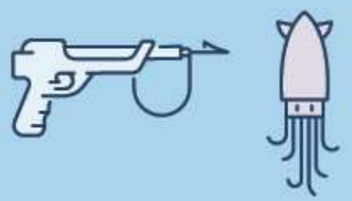
- Autorevolezza
- Senso di colpa
- Panico
- Ignoranza
- Desiderio
- Avidità
- Compassione

hing vs. spear phishing vs. whaling

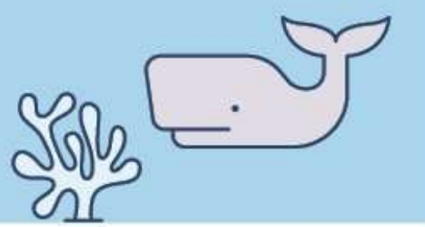
Whaling is a specific type of spear phishing, and spear phishing is a specific type of phishing. Learn the differences below.

Phishing
A broader term that covers any type of attack that tries to fool a victim into taking some action. Does not have a specific target.

Spear phishing
A type of phishing that targets individuals.



Whaling
A form of spear phishing that targets high-ranking victims within a company.



ICONS: MARINASHEVCHENKO/ADOBE STOCK, FOURLEAFLOVER/ADOBE STOCK

Alcune «confezioni» di phishing

- Grandparent scam
- Romance scam
- Cryptocurrency scam
- Employment scam
- Online account tax scam
- Lottery scam
- Charity scam
- Extended warranties scam
- Free trials scam
- … anti-scam scam

Phishing


Ministero della Giustizia

CONVOCAZIONE

Alla sua attenzione,

Su istruzioni del commissario **Troels Oerting**, direttore del Centro europeo per la criminalità informatica, vi inviamo questa citazione in conformità alle disposizioni dell'articolo 650 del codice penale. Infatti, a seguito di un sequestro in collaborazione con il Centro Nazionale per l'Analisi delle Immagini Pornografiche, Pedopornografiche e Ciberpornografiche e l'Organizzazione Internazionale di Polizia Criminale (INTERPOL), abbiamo ottenuto prove che indicano che lei ha commesso o è coinvolto in diversi reati su Internet.

I crimini che sono stati identificati e che potrebbero essere perseguiti sono i seguenti:

* **PORNOGRAFIA INFANTILE**
* **TRAFFICO DI BAMBINI**
* **ESIBIZIONISMO**

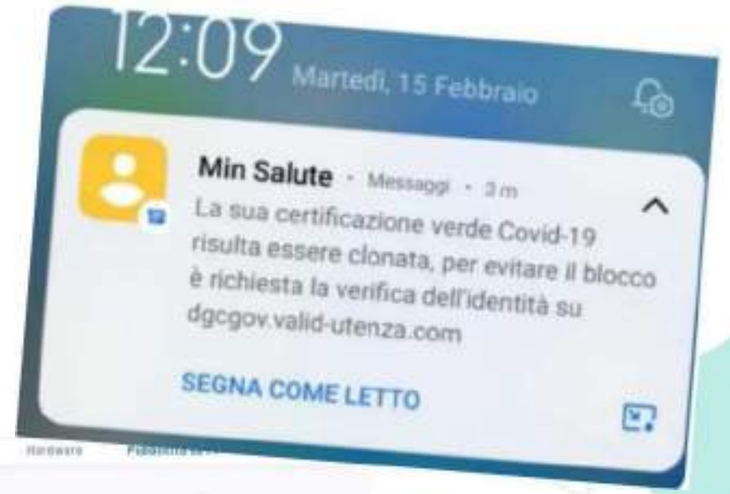
Secondo la direttiva 2011/92/ue del parlamento europeo e del consiglio del 13 dicembre 2011 sulla lotta contro l'abuso e lo sfruttamento sessuale dei minori e la pornografia infantile, questi diversi reati sono previsti e puniti dalla legge con pene **che vanno da 1 a 3 anni di reclusione con una multa di €45.000 euro**, l'iscrizione nel registro nazionale degli autori di reati sessuali e l'ampia diffusione dell'identità dell'autore del reato in tutti i media nazionali. Si prega di spiegare scrivendo all'indirizzo e-mail del Centro europeo per la criminalità informatica Ufficio nazionale EC3 | Europol, [**ec3europol.cybunit@secupyple.com**].

Dopo le 72 ore, sarà emesso un mandato d'arresto contro di lei.

Cordiali saluti,

Comandante Generale dell'Arma dei Carabinieri,
Gen. C.A. Teo Luzi

Gen. Teo Luzi



WhatsApp: phishing con messaggi vocali via email

Solo pochi giorni fa era stato annunciato un corposo aggiornamento. Purtroppo anche i cybercriminali sono sempre attivi, come dimostra la nuova campagna di **phishing** che sfrutta proprio i **messaggi vocali** di **WhatsApp**, una delle funzionalità più popolari del servizio di messaggistica. Il **malware** distribuito via email consente il furto di numerosi dati personali.

Luca Casarotti

Phishing


Fisco Oggi
RIVISTA ONLINE DELL'AGENZIA DELLE ENTRATE
Le Guide dell
Attualità
Normativa e prassi
Giurisprudenza
Dati e statistiche
Analisi e commenti
NEWS: Leonardo consegna al Qatar i primi 2 elicotteri navali NH90
Attualità
L'Agenzia lancia un nuovo allarme, tentativi di phishing ancora in azione
6 Aprile 2022
Le mail-truffa che sembrano provenire dall'amministrazione finanziaria questa volta riguardano il mancato versamento dell'imposta di bollo su fatture elettroniche
The Hacker News
Home
Data Breaches
Cyber Attacks
Vulnerabilities
Malware
Offers
Contact
Subscribe to Newsletter
New Browser-in-the Browser (BITB) Attack Makes Phishing Nearly Undetectable
March 21, 2022
Ravie Lakshmanan
Log in or sign up in seconds
Sign in
to continue to Canva
Continue with email
Popular This Week
CISA Warns of Active Exploitation of Critical Spring4Shell Vulnerability
Chinese Hackers Target VMware Horizon Servers with Log4Shell to Deploy Rootkit
15-Year-Old Bug in PEAR PHP Repository Could've Enabled Supply Chain Attacks
GitLab Releases Patch for Critical Vulnerability That Could Let Attackers Hijack Accounts
Germany Shuts Down
Google Updates from Threat Analysis Group (TAG)
THREAT ANALYSIS GROUP
Tracking cyber activity in Eastern Europe
Share

Tecniche usate nella sostituzione di persona (parente/amico in difficoltà, banca, autorità giudiziaria, agenzia delle entrate etc)

- Spoofing del caller ID
- AI
- Email e (falsi) documenti
- Chiamate vocali (possono avere informazioni che vi riguardano)

Panico – dirty laundry scam (tutti abbiamo qualcosa da nascondere…)

Hi there! I regret to inform you about some sad news for you. Approximately a month or two ago I have succeeded to gain a total access to all your devices utilized for browsing internet. Moving forward, I have started observing your internet activities on continuous basis…

Minaccia di diffusione di contenuti private

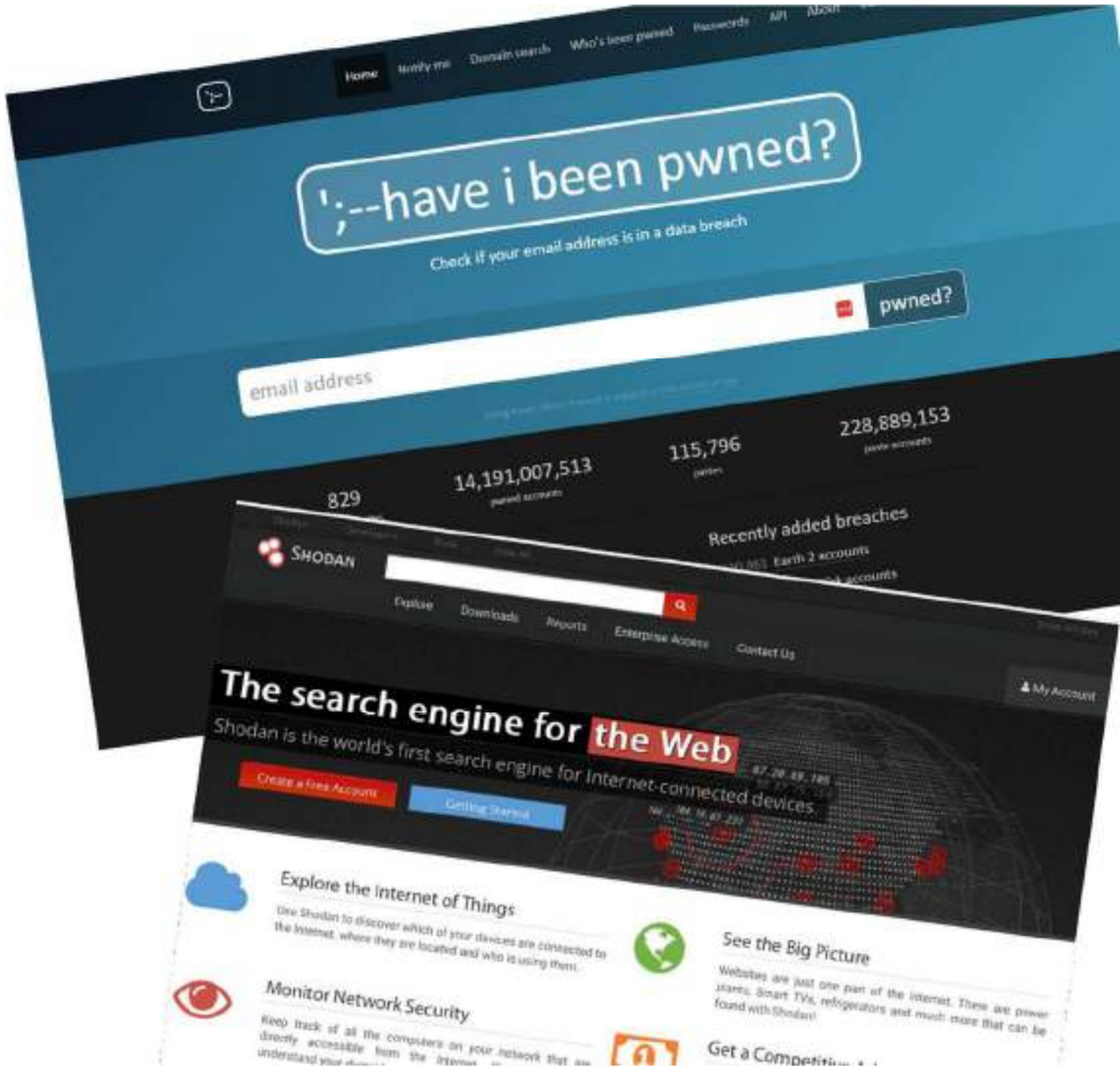
Richiesta di riscatto in bitcoin

“non contattare le forze dell’ordine”

“se entro 48 ore non paghi…”

Panico – dirty laundry scam

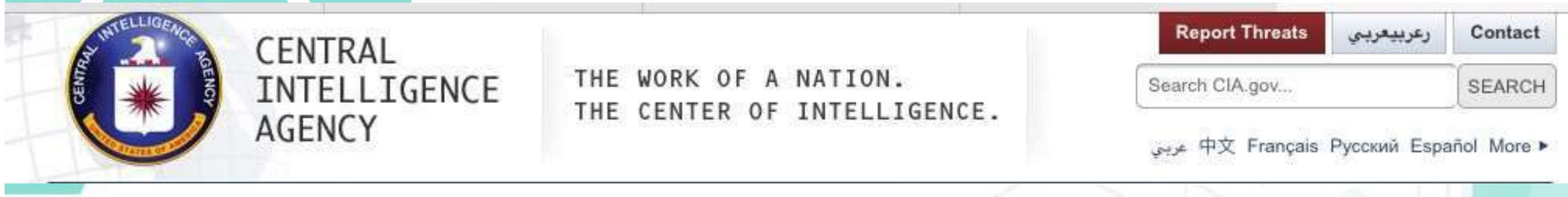
- Come può conoscere le mie password?
- Ma… sono davvero le mie password?



Chissà dove mai hanno scoperto che…

Video https://www.youtube.com/watch?v=qYnmfBiomlo

Le tecniche di OSINT



- Information does not have to be secret to be valuable. Whether in the blogs we browse, the broadcasts we watch, or the specialized journals we read, there is an endless supply of information that contributes to our understanding of the world. The Intelligence Community generally refers to this information as Open Source Intelligence (OSINT). **OSINT plays an essential role in giving the national security community as a whole insight and context at a relatively low cost.**
- OSINT is drawn from publicly available material, including:
  - The Internet
  - Traditional mass media (e.g. television, radio, newspapers, magazines)
  - Specialized journals, conference proceedings, and think tank studies
  - Photos
  - Geospatial information (e.g. maps and commercial imagery products)…

Alcune tecniche di SE: Baiting (esca)

Il "baiting" è una tecnica di social engineering che sfrutta la curiosità o la grettezza dell'utente per indurlo a eseguire una certa azione. Si basa sulla creazione di un'esca (da qui il nome "baiting") per attirare l'utente.

Un esempio comune di baiting potrebbe essere un attaccante che lascia una chiavetta USB in un luogo pubblico, come un parcheggio o una caffetteria. La chiavetta USB potrebbe avere un'etichetta che indica contenuti interessanti o preziosi, come "Stipendi dei dipendenti" o "Esame finale con risposte". Se una persona trova la chiavetta USB e la inserisce nel proprio computer per vedere cosa contiene, potrebbe involontariamente installare malware o consentire all'attaccante di accedere al proprio sistema.

Online, il baiting (clickbaiting) può assumere la forma di annunci o messaggi che promettono premi, film gratuiti, software o altri contenuti desiderabili, ma che in realtà portano a siti web infetti o al download di file dannosi.

Non fidatevi dei dispositivi «smarriti»

Users Really Do Plug in USB Drives They Find

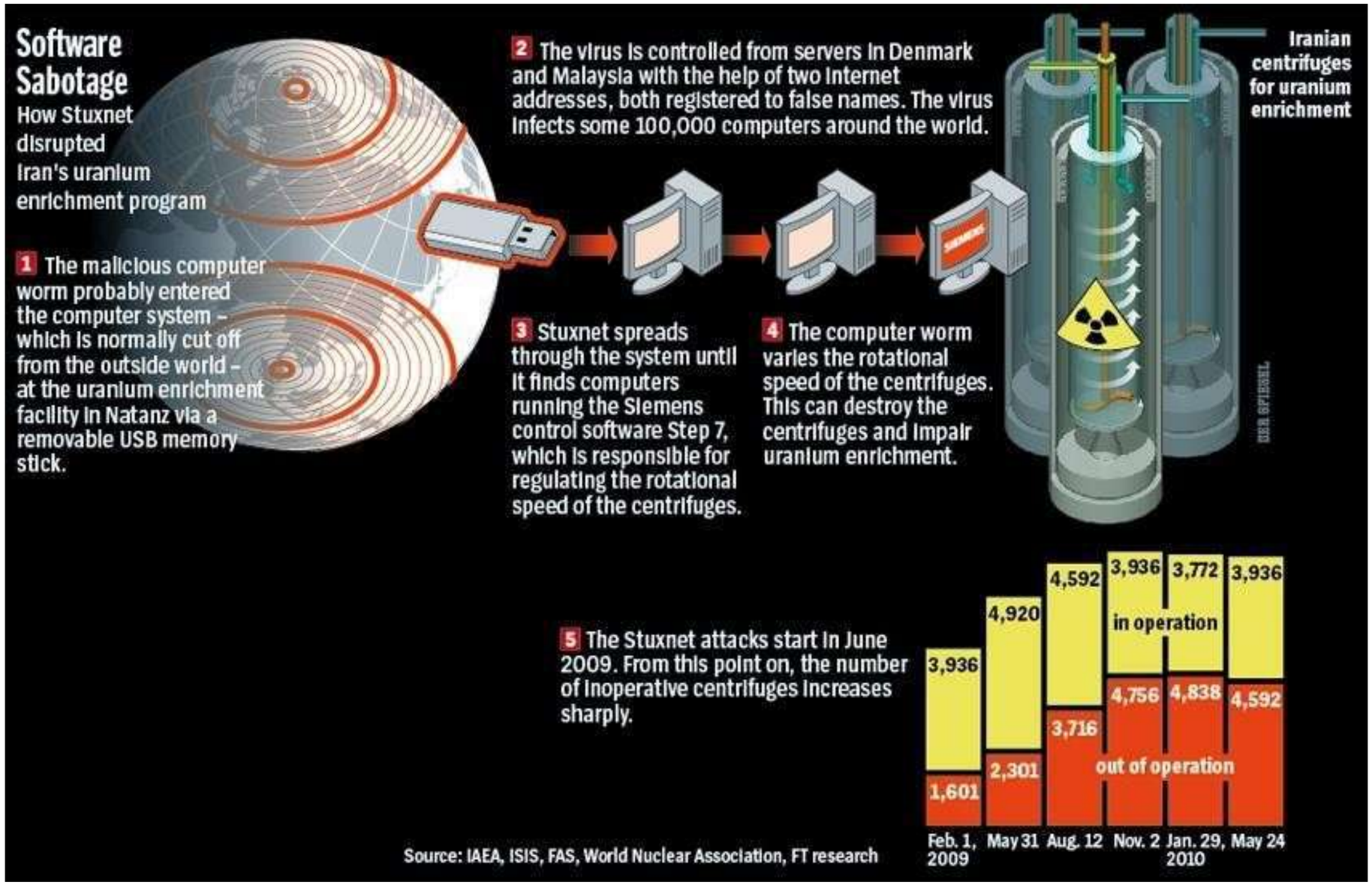
Matthew Tischer† Zakir Durumeric†‡ Sam Foster† Sunny Duan†
Alec Mori† Elie Bursztein◇ Michael Bailey†

† University of Illinois, Urbana Champaign ‡ University of Michigan ◇ Google, Inc.
{tischer1, sfoster3, syduan2, ajmori2, mdbailey}@illinois.edu
zakir@umich.edu elieb@google.com

Abstract—We investigate the anecdotal belief that end users will pick up and plug in USB flash drives they find by completing a controlled experiment in which we drop 297 flash drives on a large university campus. We find that the attack is effective with an estimated success rate of 45–98% and expeditious with the first drive connected in less than six minutes. We analyze the types of drives users connected and survey those users to understand their motivation and security profile. We find that

median time to connection of 6.9 hours and the first connection
occurring within six minutes from when the drive was dropped
Contrary to popular belief, the appearance of a drive does no
increase the likelihood that someone will connect it to their
computer. Instead, users connect all types of drives unless
there are other means of locating the owner—suggesting tha
participants are altruistically motivated. However, while use

Francesco Paolo Micozzi


Software Sabotage
How Stuxnet disrupted Iran's uranium enrichment program
1 The malicious computer worm probably entered the computer system – which is normally cut off from the outside world – at the uranium enrichment facility in Natanz via a removable USB memory stick.
2 The virus is controlled from servers in Denmark and Malaysia with the help of two Internet addresses, both registered to false names. The virus infects some 100,000 computers around the world.
3 Stuxnet spreads through the system until it finds computers running the Siemens control software Step 7, which is responsible for regulating the rotational speed of the centrifuges.
4 The computer worm varies the rotational speed of the centrifuges. This can destroy the centrifuges and impair uranium enrichment.
Iranian centrifuges for uranium enrichment
5 The Stuxnet attacks start in June 2009. From this point on, the number of inoperative centrifuges increases sharply.
in operation
out of operation
3,936
1,601
4,920
2,301
4,592
3,716
3,936
4,756
3,772
4,838
3,936
4,592
Feb. 1, 2009
May 31
Aug. 12
Nov. 2
Jan. 29, 2010
May 24
Source: IAEA, ISIS, FAS, World Nuclear Association, FT research

Alcune tecniche di SE: Pretexting

Il pretexting è una tecnica di ingegneria sociale in cui un attaccante crea una falsa storia (pretesto) per convincere una vittima a rilasciare informazioni o ad eseguire determinate azioni. Questa tecnica si basa su inganni e manipolazioni psicologiche per ottenere accesso a informazioni sensibili o riservate.

Un esempio comune di pretexting è quando un attaccante si finge un tecnico del servizio clienti, un agente di polizia, o un altro individuo in una posizione di autorità o di fiducia. L'attaccante può poi convincere la vittima a rilasciare informazioni sensibili, come password o numeri di conto bancario, o a eseguire azioni che compromettono la sicurezza, come installare software dannoso o accedere a link dannosi.

Il pretexting è un metodo potente e pericoloso di ingegneria sociale perché sfrutta la tendenza naturale delle persone a fidarsi di figure di autorità o di persone che sembrano avere buone intenzioni.

Alcune tecniche di SE: Pretexting

Pretexting attraverso il telefono: Un attaccante potrebbe chiamare un dipendente di un'organizzazione affermando di far parte del reparto IT. Potrebbe spiegare che c'è un problema con l'account dell'utente e che ha bisogno delle credenziali dell'utente per risolverlo. L'attaccante potrebbe anche fornire un numero di identificazione del dipendente falso o altre informazioni per sembrare legittimo. Se il dipendente fornisce le credenziali, l'attaccante può quindi accedere al sistema come se fosse l'utente.

Alcune tecniche di SE: Esempio di Pretexting

- Un truffatore potrebbe chiamare fingendosi il supporto tecnico del tuo computer/smartphone, sostenendo di aver rilevato problemi di sicurezza. Chiederà di installare software o fornire accesso remoto "per risolvere il problema".
- Qualcuno potrebbe chiamare fingendosi la tua banca, dicendo che c'è un problema con il conto o la carta. Chiederà di verificare dati personali o effettuare operazioni "di emergenza".
- Il truffatore finge di essere un parente in difficoltà che ha bisogno urgentemente di denaro.

Alcune tecniche di SE: Pretexting

Pretexting via email (Phishing): In questo esempio, un attaccante potrebbe inviare un'email che sembra provenire da un'organizzazione affidabile, come una banca o un fornitore di servizi Internet. L'email potrebbe affermare che c'è un problema con l'account dell'utente e che l'utente deve cliccare su un link e inserire le sue credenziali per risolverlo. Se l'utente clicca sul link e inserisce le sue credenziali, l'attaccante può accedere all'account dell'utente.

Alcune tecniche di SE: Quid Pro Quo, Diversion Theft, Dumpster Diving e Shoulder Surfing

Quid Pro Quo Attacks: In questo caso, l'attaccante offre qualcosa in cambio di informazioni. Ad esempio, potrebbe offrire supporto tecnico gratuito in cambio della password di un utente.

Diversion Theft: L'attaccante devia la consegna di merci o lettere a un altro indirizzo per ottenere accesso a informazioni o beni.

Dumpster Diving: Gli attaccanti cercano documenti sensibili nei rifiuti.

Shoulder Surfing: Gli attaccanti osservano direttamente la vittima inserire informazioni sensibili, come una password, su una tastiera.

Alcune tecniche di SE: Shouldersurfing

Intelligenza artificiale e attacchi informatici

IA e sicurezza informatica

Automazione degli attacchi: I ciber-attaccanti possono utilizzare l'IA per automatizzare e velocizzare gli attacchi. Ad esempio, possono usare l'IA per condurre attacchi di forza bruta più efficienti, o per scoprire e sfruttare vulnerabilità nei sistemi di sicurezza.

Phishing e social engineering: L'IA può essere utilizzata per creare messaggi di phishing più convincenti o per condurre attacchi di ingegneria sociale. Ad esempio, un attaccante potrebbe utilizzare l'IA per generare e-mail di phishing che sembrano provenire da un contatto fidato, o per impersonare una persona reale in una chat o in una conversazione telefonica.

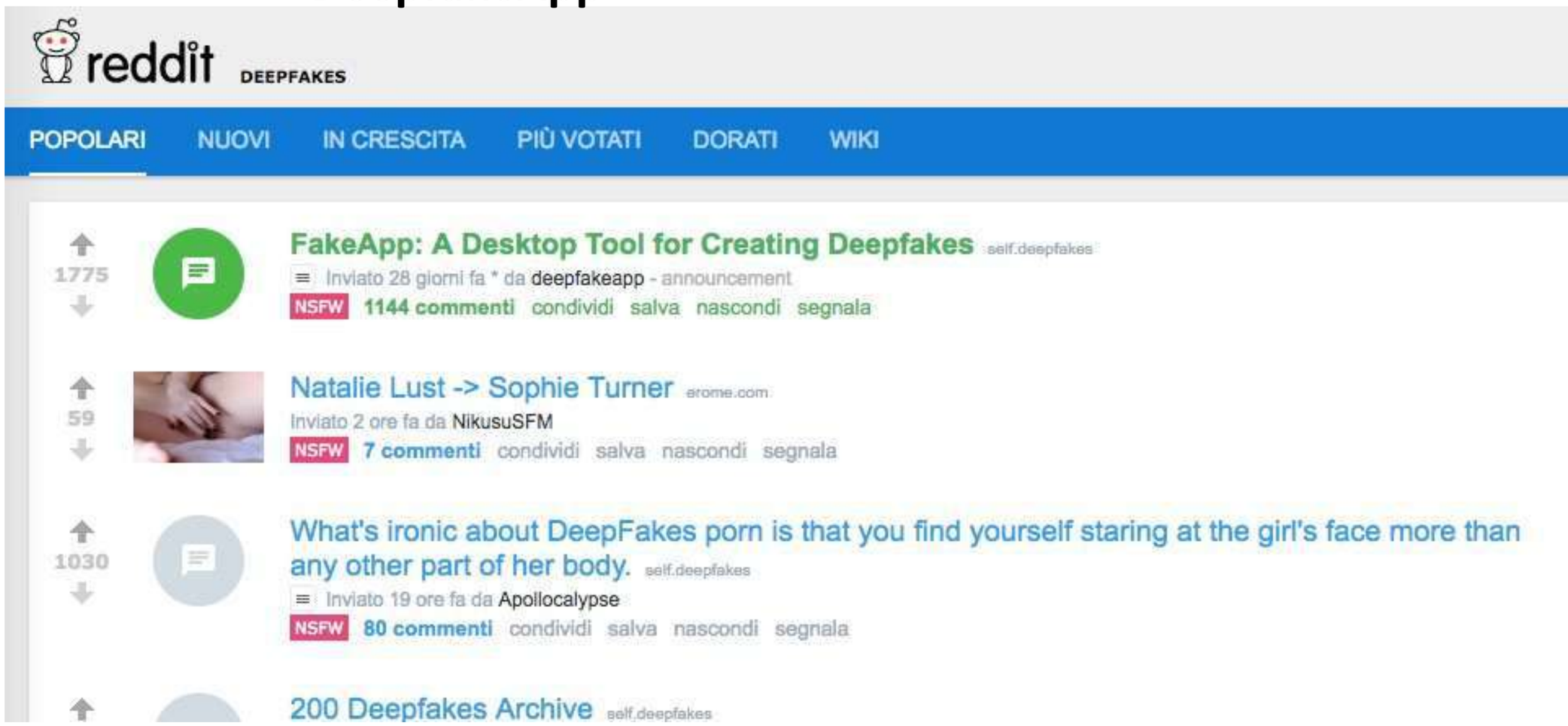
Malware AI-driven: L'IA può essere utilizzata per creare malware più sofisticati e difficili da rilevare. Ad esempio, un attaccante potrebbe utilizzare l'IA per creare un malware che può adattarsi e cambiare il suo comportamento per evitare la rilevazione da parte dei sistemi di sicurezza.

Attacchi AI contro AI: Gli attaccanti possono utilizzare l'IA per sfruttare le debolezze negli algoritmi di apprendimento automatico. Ad esempio, potrebbero utilizzare tecniche di attacco come l'adversarial machine learning per ingannare un sistema di IA e farlo comportare in modo indesiderato.

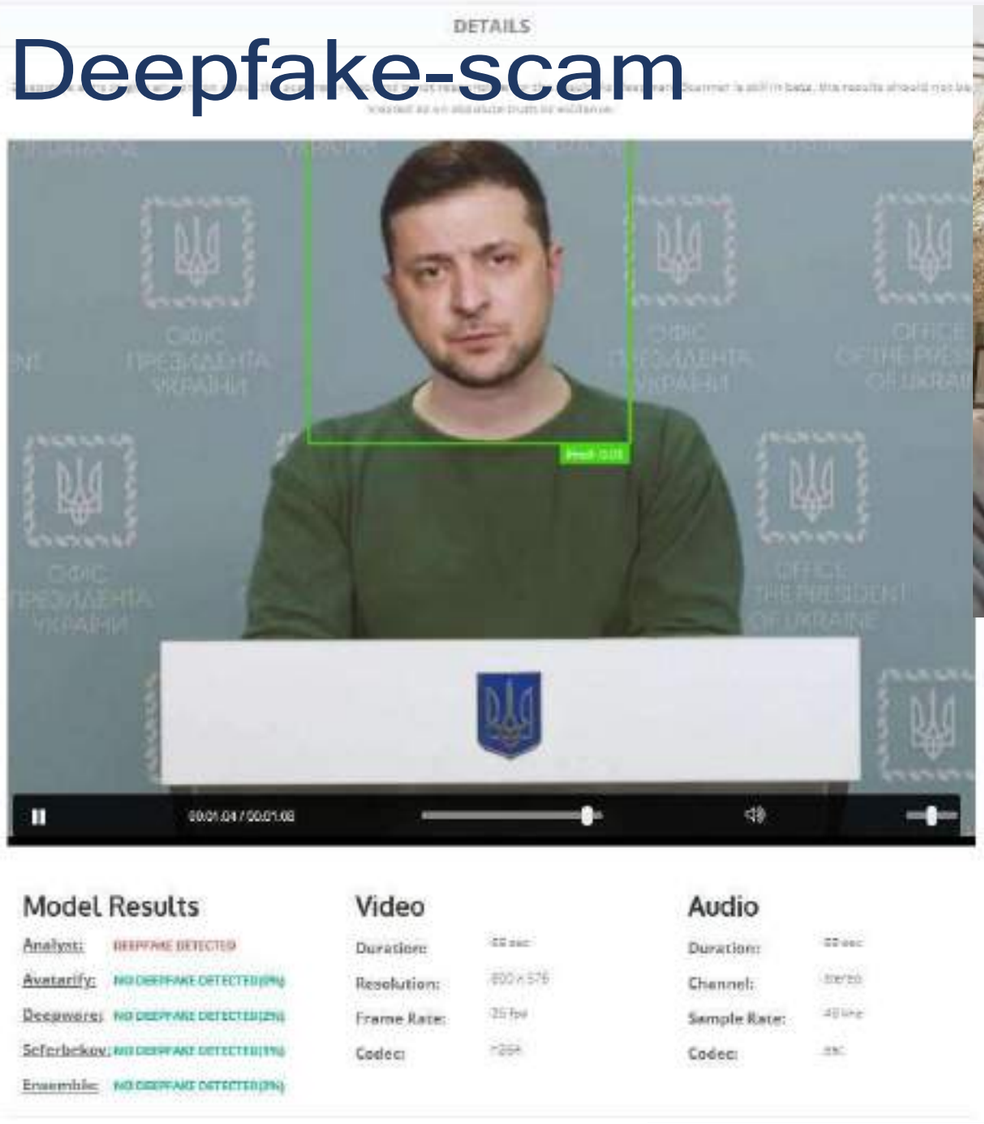
Abuso di servizi basati su IA: Gli attaccanti possono abusare dei servizi basati su IA per condurre attacchi. Ad esempio, potrebbero utilizzare servizi di generazione di testo basati su IA per creare contenuti dannosi o ingannevoli

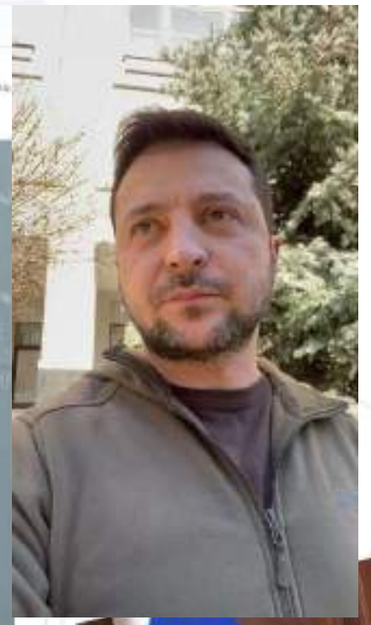
Link video: https://www.youtube.com/watch?v=ohmajJTcpNk

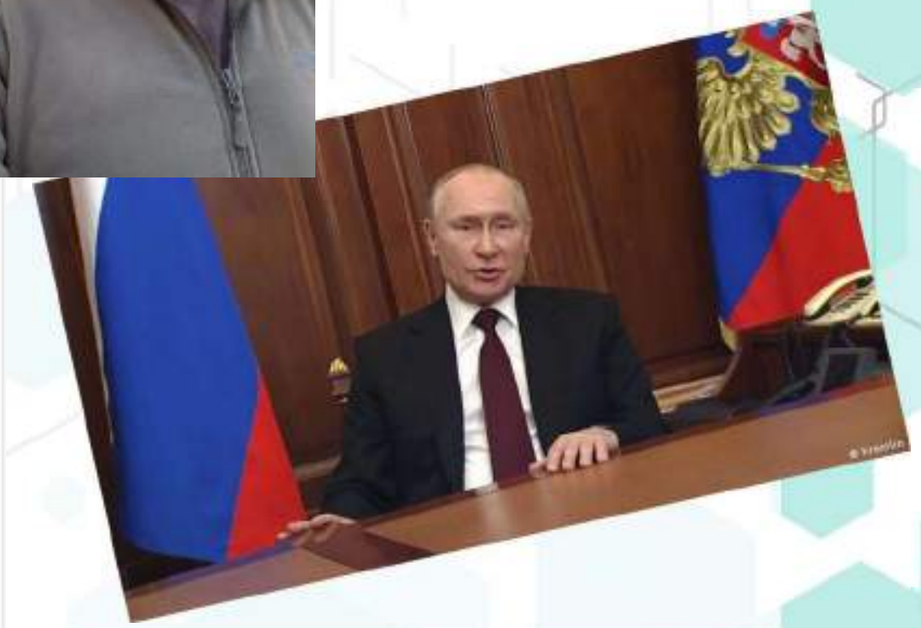
DeepfakeApp



Deepfake-scam







Deepfake-scam

https://thispersondoesnotexist.com/

Diffidate, gente. Diffidate!

Grazie per l’attenzione!