La Cybersecurity dei sistemi intelligenti: dai report dell'ENISA al Regolamento "AI Act"

WOMEN for Security

Parola di Cyber Ladies

Avv. Anna Capoluongo

La Cybersecurity dei sistemi intelligenti: dai report dell'ENISA al Regolamento "AI Act"

Avv. Anna Capoluongo[1]

Sommario:

Il contributo si propone di approcciare alla tematica della cybersecurity attraverso un *focus* mirato sulle norme in materia contenute nei report dell'ENISA, del JCR e nella versione (quasi) consolidata del testo dell'AI Act, così come approvato – ad oggi - dal Consiglio e dal Parlamento europei, per chiudere con un breve accenno al Cyber Resilience Act.

Parole Chiave: AI Act – cybersecurity – data protection – risk-based approach – cyber resilience

I. Introduzione

Per poter meglio approfondire le tematiche oggetto del presente contributo, giova, anzitutto, ricordare che con il termine *cybersecurity* (anche cyber sicurezza o sicurezza informatica) si vuole far riferimento a quell'insieme di tecnologie, processi e misure di protezione progettate per ridurre il rischio di attacchi informatici. Una serie di azioni, quindi, volte alla protezione (anticipata) e alla difesa (anche successiva) e resilienza di sistemi elettronici, reti, server e dispositivi, in ottica di sicurezza informatica e delle informazioni.
Se è vero che si è soliti utilizzare questo termine in chiave omnicomprensiva, è altrettanto importante evidenziare come, in realtà, della cyber sicurezza facciano parte differenti "categorie", che si vogliono qui sinteticamente elencare, per maggiore chiarezza e comprensibilità del contributo:

- sicurezza delle reti informatiche;
- sicurezza delle applicazioni (software e dispositivi);
- sicurezza delle informazioni (comprensive di dati personali);
- sicurezza delle applicazioni;
- sicurezza operativa degli asset;
- disaster recovery e business continuity,

senza dimenticare un elemento di grandissima importanza (qui come nel panorama della *data protection* e della responsabilità nell'uso dei sistemi intelligenti), ossia la formazione e sensibilizzazione degli operatori/utenti/utilizzatori (a cui, si noti, nell'AI Act si fa attualmente riferimento mediante il termine "alfabetizzazione").

Ciò premesso, l'obiettivo principe della *cybersecurity*, come anticipato, è quello di proteggere il perimetro di sicurezza da minacce ed attacchi, che vanno, quindi, debitamente compresi e distinti per poter essere affrontati a dovere.
Ne *L'Arte della Guerra* si leggeva "*Conosci il nemico, conosci te stesso, mai sarà in dubbio il risultato di 100 battaglie*", e – *mutatis mutandis* – si può certamente applicare anche in questo contesto, laddove i principali "nemici" prendono il nome di:

- *cybercrime,* che si definisce come quel "*reato nel quale la condotta o l'oggetto materiale del crimine sono correlati a un sistema informatico o telematico, ovvero perpetrato utilizzando un tale sistema o colpendolo (rispettivamente, si parla di computer as a tool e computer as a target)*"[2],

[1] Avvocato Data Protection & ICT| Data Protection Officer UNI 11697 | membro dell'EDPB's Support Pool of Experts| Membro Comitato di Presidenza E.N.I.A. (Ente Nazionale per l'AI) | Membro Women For Security | Membro Osservatorio sulla Giustizia Civile Tribunale Milano (Gruppo sul danno da illecito trattamento dei dati personali) | Afferente B-ASC (Università Milano-Bicocca Applied Statistics Center)| Vicepresidente Institute for Research of Law Economical and Social Studies| Docente a c. (Università di Padova, Sole24Ore Business School).
[2] https://www.treccani.it/enciclopedia/cybercrime_(Lessico-del-XXI-Secolo)/#.

- *cyberterrorismo* (utilizzo di tecnologie informatiche al fine di sviluppare un'azione o una strategia terroristica[3]), e
- *cyberattacchi,* ossia tentativi di ottenere l'accesso non autorizzato ai sistemi informatici al fine di appropriarsi, modificare o distruggere dati. E così, una qualunque manovra che colpisca sistemi informatici, infrastrutture, reti di calcolatori e/o dispositivi elettronici, tramite atti malevoli, finalizzati al furto, alterazione o distruzione di specifici obiettivi violando sistemi suscettibili[4].
  Le tipologie di cyberattacchi sono innumerevoli e fanno leva principalmente su 3 fattori: la paura, la vulnerabilità e il clamore che deriva dalla notizia di un avvenuto attacco. È, infine, opportuno sapere che:
  - si possono dividere in due categorie, *attacchi sintattici* (diretti, consistono nella diffusione di malware e software malevolo) e *semantici* (indiretti, consistono nella modifica di informazioni corrette e nella diffusione di informazioni errate);
  - *ex multis*, i più noti sono quelli che hanno ad oggetto database di informazioni e password (attacco a dizionario e *brute force attack*), il cd. *man in the middle* (intercettazioni di conversazioni e sottrazione dati), il *denial of service* (sovraccarico dei server con traffico in eccesso), l'immissione di codice SQL (Structured Language Query), il *phishing*, il *social engineering* e i *malware* o software malevoli (in cui rientrano gli ormai - tristemente - noti *ransomware*).

Sempre sul punto, si ritiene di certo interesse evidenziare come l'ENISA (l'Agenzia dell'Unione Europea per la Cybersicurezza) abbia pubblicato[5] di recente la sintesi esecutiva di quest'anno "*Foresight Cybersecurity Threats for 2030*", presentando una panoramica della *top ten* delle minacce *cybersecurity*, che include:
- due *new entry*, quali lo sfruttamento di sistemi non patchati e non aggiornati all'interno di un ecosistema tecnologico intersettoriale sovraccarico, e l'impatto fisico delle interruzioni naturali/ambientali sulle infrastrutture digitali critiche, oltre a
- compromissione della catena di fornitura delle dipendenze del software;
- carenza di competenze;
- errore umano e sistemi legacy sfruttati all'interno di ecosistemi cyber-fisici;
- aumento dell'autoritarismo della sorveglianza digitale / perdita della privacy;
- fornitori di servizi ICT transfrontalieri come singolo punto di fallimento;
- campagne di disinformazione avanzata / operazioni di influenza (IO);
- aumento delle minacce ibride avanzate;
- abuso dell'intelligenza artificiale.

Tutto ciò doverosamente premesso, i paragrafi che seguono si prefiggono lo scopo di fornire una panoramica attuale – senza pretesa di piena esaustività – degli aspetti più rilevanti della materia della *cybersecurity* nello specifico ambito dei sistemi di intelligenza artificiale (AI), profilo questo affrontato a più livelli, partendo dalle riflessioni dell'ENISA del 2023, passando per il report del JRC[6] (Centro Comune di Ricerca UE) e arrivando, infine, alle norme specificamente previste in tal senso nel Regolamento sull'Intelligenza Artificiale (AI Act), prossimo alla pubblicazione in Gazzetta Ufficiale.

II. Il report dell'ENISA sulla cybersicurezza dell'Artificial Intelligence[7]

Il report dell'ENISA del marzo 2023 apre sottolineando, correttamente, come comprendere l'intelligenza artificiale e la sua portata sia il primo passo verso la definizione della sicurezza informatica (*cybersecurity*), anche se una

[3] https://www.treccani.it/enciclopedia/cyberterrorismo_%28Lessico-del-XXI-Secolo%29/.
[4] https://it.wikipedia.org/wiki/Attacco_informatico.
[5] https://www.enisa.europa.eu/news/skills-shortage-and-unpatched-systems-soar-to-high-ranking-2030-cyber-threats.
[6] Il JRC fornisce competenze e conoscenze scientifiche indipendenti e basate su dati concreti, sostenendo le politiche dell'UE per avere un impatto positivo sulla società: https://commission.europa.eu/about-european-commission/departments-and-executive-agencies/joint-research-centre_it.
[7] https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation.

definizione e un ambito di applicazione chiari di tale concetto paiano quanto meno sfuggenti. In tal senso, come vedremo, la pubblicazione della prima normazione in tema di AI (l'AI Act europeo, appunto), potrà venire in aiuto, poiché propone – finalmente – una definizione sufficientemente univoca di intelligenza artificiale, sino ad oggi non presente[8].
L'Agenzia, inoltre, rimandando ad un proprio precedente contributo del 2021[9], si sofferma sul richiamare sinteticamente la relazione multidimensionale tra AI e *cybersecurity*, identificando le seguenti tre dimensioni:

1. la *cybersecurity* dell'IA, con *focus* primario sulla mancanza di robustezza e vulnerabilità dei modelli e degli algoritmi di AI,
2. l'AI a supporto della *cybersecurity*, da utilizzarsi come strumento/mezzo per creare una cybersecurity avanzata e per facilitare gli sforzi delle forze dell'ordine e di altre autorità pubbliche per rispondere meglio alla criminalità informatica,
3. l'uso malevolo dell'AI, per creare tipologie di attacchi sempre più sofisticati.

Circa la prima dimensione, è giusto il caso di riflettere sul fatto che le interpretazioni di *cybersecurity* dell'AI da prendere in considerazione possono essere differenti, partendo da un approccio ristretto e tradizionale (protezione contro gli attacchi alla riservatezza, all'integrità e alla disponibilità degli asset durante il ciclo di vita di un sistema di IA) - su cui si concentra il report del 2023 in oggetto-, per arrivare sino a prospettive più estese, che ricomprendano l'affidabilità dei sistemi, la qualità dei dati, la supervisione, la robustezza, l'accuratezza, la spiegabilità, la trasparenza e la tracciabilità.
È evidente che i legami tra *cybersecurity* e affidabilità siano assai complessi e che i requisiti di affidabilità integrino e talvolta si sovrappongano a quelli della *cybersecurity* dell'AI nel garantirne il corretto funzionamento.

Il Report dell'ENISA si snoda, successivamente, nei capitoli 3 e 4, rispettivamente dedicati al concetto di standardizzazione a supporto della *cybersecurity* e all'analisi sulla della copertura degli standard più rilevanti rispetto al modello di sicurezza CIA[10] e alle caratteristiche di affidabilità a supporto della *cybersecurity,* per poi chiudere con delle brevi conclusioni e un allegato A sugli standard applicabili alla materia in oggetto.
Quanto al capitolo 3, l'attenzione si concentra principalmente sui soli standard che possono essere armonizzati, limitando, quindi, l'analisi a quelli dell'Organizzazione Internazionale per la Standardizzazione (ISO) e della Commissione Elettrotecnica Internazionale (IEC), del Comitato Europeo per la Standardizzazione (CEN) e del Comitato Europeo per la Standardizzazione Elettrotecnica (CENELEC), e dell'Istituto Europeo per gli Standard di Telecomunicazione (ETSI).
Il capitolo 4, invece, dopo aver riassunto l'applicazione del paradigma CIA nel contesto dell'AI nella seguente tabella[11]

[8] Per approfondimenti si veda A. Capoluongo, "*Smart cities tra intelligenza artificiale, videosorveglianza e data protection*", 2023, ed. Simone Professionale.
[9] Il rapporto ENISA Securing Machine Learning Algorithms, https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.
[10] O "RID", riservatezza, integrità e disponibilità.
[11] Per approfondimenti, si vedano anche: White Paper 'Towards auditable AI systems' of Germany's Federal Office for Information Security (https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf?__blob=publicationFile&v=6); ENISA report Securing Machine Learning Algorithms (https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms).

Security goal	Contextualisation in AI (selected examples of AI-specific attacks)
Confidentiality	**Model and data stealing attacks:** **Oracle:** A type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be precursor steps to more harmful types, for example evasion or poisoning. It is as if the attacker made the model talk to then better compromise it or to obtain information about it (e.g. model extraction) or its training data (e.g. membership inference attacks and inversion attacks). **Model disclosure:** This threat refers to a leak of the internals (i.e. parameter values) of the ML model. This model leakage could occur because of human error or a third party with too low a security level.
Integrity	**Evasion:** A type of attack in which the attacker works on the ML algorithm's inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors). It is as if the attacker created an 'optical illusion for the algorithm. Such modified inputs are often called adversarial examples. **Poisoning:** A type of attack in which the attacker alters data or models to modify the ML algorithm's behaviour in a chosen direction (e.g. to sabotage its results or to insert a back door). It is as if the attacker conditioned the algorithm according to its motivation.
Availability	**Denial of service:** ML algorithms usually consider input data in a defined format to make their predictions. Thus, a denial of service could be caused by input data whose format is inappropriate. However, it may also happen that a malicious user of the model constructs an input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service.

Application of CIA paradigm in the context of AI, Cybersecurity of AI and Standardisation Report (ENISA)

si sofferma ad indicare[12], quali standard particolarmente rilevanti i seguenti:
"*ISO/IEC 27001, Information security management, and ISO/IEC 27002, Information security controls: relevant to all security objectives,*
ISO/IEC 9001, Quality management system: especially relevant to integrity (e.g. in particular for data quality management to protect against poisoning) and availability".

Il report si conclude con una serie di raccomandazioni rivolte a:

- tutte le organizzazioni: utilizzare una terminologia standardizzata e armonizzata per la sicurezza informatica, comprese le caratteristiche di affidabilità e una tassonomia dei diversi tipi di attacchi specifici ai sistemi di AI;
- alle organizzazioni che sviluppano standard:
    - sviluppare orientamenti specifici/tecnici su come applicare all'intelligenza artificiale gli standard esistenti relativi alla sicurezza informatica del software,
    - caratteristiche intrinseche del machine learning dovrebbero riflettersi negli standard,
    - garantire che siano stabiliti collegamenti tra i comitati tecnici sulla sicurezza informatica e i comitati tecnici sull'intelligenza artificiale in modo che gli standard dell'intelligenza artificiale sulle caratteristiche di affidabilità (supervisione, robustezza, accuratezza, spiegabilità, trasparenza, ecc.) e sulla qualità dei dati includano potenziali problemi di sicurezza informatica;
- e, in generale, in vista dell'attuazione dell'AI Act:
    - data l'applicabilità dell'IA in un'ampia gamma di settori, l'identificazione dei rischi per la sicurezza informatica e la determinazione di requisiti di sicurezza adeguati dovrebbero basarsi su un'analisi specifica del sistema e, ove necessario, su standard specifici del settore;
    - incoraggiare la ricerca e lo sviluppo nei settori in cui la standardizzazione è limitata dallo sviluppo tecnologico;
    - sostenere lo sviluppo di standard per gli strumenti e le competenze degli attori che effettuano la valutazione della conformità;
    - garantire la coerenza tra AI Act e altre iniziative legislative sulla cybersicurezza.

[12] Se si considerano i sistemi di IA come software e si prendono in considerazione il loro intero ciclo di vita, gli standard generali, cioè quelli che non sono specifici per l'IA e che riguardano aspetti tecnici e organizzativi, possono contribuire a mitigare molti dei rischi che l'IA deve affrontare.

III. Il report del JRC (Joint Research Centre, European Commission): *Cybersecurity of Artificial Intelligence in the AI Act*[13]

Il rapporto in oggetto si concentra sull'individuazione di principi guida per affrontare il requisito di sicurezza informatica per i sistemi AI ad alto rischio, quindi con espresso riferimento all'articolo 15 della proposta della Commissione europea per la legge sull'AI (AI Act).
È bene sottolineare come il rapporto sia stato pubblicato a settembre 2023, e quindi facente riferimento alla versione della proposta di Regolamento in allora disponibile. Nel capitolo seguente, il presente contributo si concentrerà sulle norme relative alla *cybersecurity* così come presenti nella versione (quasi) consolidata del testo, come approvato – ad oggi - dal Parlamento europeo[14].
Pare, in tal senso, opportuno ribadire che qualsiasi valutazione o commento nel contesto del presente contributo si riferisce ad una versione ancora non approvata del testo, che solo una volta divenuto definitivo (con pubblicazione in Gazzetta Ufficiale europea) entrerà in vigore. Da quel momento, sono comunque previsti altri 24 mesi per la piena attuazione, anche se:

- alcune norme saranno operative dopo sei mesi (ad esempio, quelle sui sistemi di AI vietati);
- alcune dopo un anno (ad esempio, *governance* dell'AI con finalità generali);
- altre dopo tre (AI ad alto rischio definiti nell'allegato II).

Spostando nuovamente il *focus* sul report del JRC, lo stesso evidenzia come il requisito di sicurezza informatica dell'AI Act si applichi al sistema AI nel suo complesso e non direttamente ai suoi componenti interni, di talché, per garantire la conformità, sarà opportuno applicare un approccio olistico, integrato e continuo, e condurre una valutazione dei rischi che tenga conto della progettazione dell'intero sistema.
Quello appena esposto, seppur brevemente, è uno dei 4 principi guida che emergono nei risultati dell'analisi del JCR, insieme ai seguenti:

- la conformità all'AI Act richiede necessariamente una valutazione del rischio di sicurezza, quindi per garantire che un sistema di IA sia *compliant* con i requisiti di *cybersecurity* previsti, è necessario condurre una valutazione del rischio di sicurezza considerando l'architettura interna del sistema di IA e il contesto applicativo previsto. Questa valutazione del rischio di *cybersecurity*, effettuata nel contesto del sistema di gestione del rischio descritto nell'articolo 9 del Regolamento, mira a identificare i rischi specifici, a tradurre i requisiti di cybersecurity di livello superiore della normativa in requisiti specifici per i componenti del sistema e a implementare le misure di mitigazione necessarie;
- la protezione dei sistemi di IA richiede un approccio integrato e continuo che utilizzi pratiche comprovate e controlli specifici per l'AI, passando per una combinazione di controlli esistenti per i sistemi software e misure specifiche per i modelli intelligenti, la cui peculiarità è di essere la somma di tutti i loro componenti e delle loro interazioni. Per garantire che i sistemi di AI siano conformi ai requisiti di *cybersecurity* dell'AI Act, è necessario adottare un approccio olistico che segua i principi di security-in-depth[15] e security-by-design[16].
- lo stato dell'arte della sicurezza dei modelli di AI presenta dei limiti. Nell'attuale panorama dell'IA coesiste un'ampia gamma di tecnologie di AI con diversi gradi di maturità. Non tutte le tecnologie di IA potrebbero essere pronte per essere utilizzate nei sistemi di IA progettati per essere impiegati in scenari ad alto rischio, a meno che non si affrontino adeguatamente i loro limiti in termini di *cybersecurity*. In alcuni casi, soprattutto

---

[13] European Commission, Joint Research Centre, Junklewitz, H., Hamon, R., André, A. et al., *Cybersecurity of artificial intelligence in the AI Act – Guiding principles to address the cybersecurity requirement for high-riskAI systems*, Publications Office of the European Union, 2023, **https://data.europa.eu/doi/10.2760/271009.**
[14] https://www.europarl.europa.eu/doceo/document/A-9-2023-0188-AM-808-808_IT.pdf.
[15] O DiD (difesa in profondità), consiste in una stratificazione delle risorse informatiche di protezione (più sistemi di sicurezza su differenti livelli), al fine di rallentare l'attacco in corso, così da mettere in atto strategie difensive più efficaci.
[16] Ossia la progettazione di un software espressamente al fine di essere sicuro, anticipando e minimizzando a priori gli impatti delle vulnerabilità che potrà manifestare in produzione.

per le tecnologie emergenti di IA, esistono limitazioni intrinseche che non possono essere affrontate esclusivamente a livello di modello di IA. In questi casi, la conformità ai requisiti di cybersicurezza della legge sull'IA può essere raggiunta solo seguendo l'approccio olistico descritto in precedenza.

La relazione evidenzia, inoltre, come i requisiti di sicurezza informatica, accuratezza e robustezza siano collegati alla dimensione tecnica dei sistemi di intelligenza artificiale, richiedendo, quindi, una profonda comprensione del funzionamento interno dei sistemi stessi, delle pratiche tecniche e degli standard consolidati, specialmente a causa delle sfide tecnologiche specifiche ed intrinseche dell'intelligenza artificiale/apprendimento automatico[17]. Queste sono principalmente da rinvenirsi nella capacità di ragionamento e apprendimento, nella forte dipendenza dai dati e nella natura stocastica (probabilistica) dei dati e possono essere raggruppate approssimativamente in due categorie:

- organizzative legate ai processi (ad es. armonizzazione di terminologie, tassonomie delle minacce e definizioni in tutti i campi e negli standard; gestire la sicurezza del ciclo di vita dell'intelligenza artificiale e i problemi di sicurezza della catena di fornitura specifici dell'intelligenza artificiale; adattare i controlli di sicurezza esistenti per il software di intelligenza artificiale);
- di ricerca e sviluppo relative alle tecniche (ad es. valutazione degli attacchi ai modelli di machine learning; sviluppare misure di sicurezza specifiche per l'intelligenza artificiale e modelli di rafforzamento per metodologie di intelligenza artificiale più avanzate; definire metriche e misure per la sicurezza informatica dell'AI e la robustezza antagonista dei modelli di intelligenza artificiale; valutare i compromessi con altri requisiti, ad esempio tra accuratezza e sicurezza informatica; sviluppare un'esperienza pratica di modellazione delle minacce dell'AI).

Sullo specifico punto della *cybersecurity* nell'AI Act, poi, il report rileva come la stessa sia disciplinata dall'articolo 15, anche se non come requisito a sé stante, ma insieme ai principi di accuratezza e robustezza, e ulteriormente approfondita nel considerando 51. In tal senso, per i sistemi di AI ad alto rischio bisognerà quindi:

- progettarli affinché siano resilienti contro i tentativi di alterarne l'uso, il comportamento e le prestazioni e per comprometterne l'uso, il comportamento e le prestazioni e di comprometterne le proprietà di sicurezza da parte di terzi malintenzionati;
- implementare soluzioni organizzative e tecniche per raggiungere tali obiettivi;
- effettuare una valutazione del rischio di sicurezza informatica;
- adeguare le soluzioni tecniche alle circostanze e ai rischi rilevanti.

Un punto d'appoggio forte per avvicinarsi a soluzioni valide in tema di *cybersecurity* per i sistemi intelligenti viene, sostanzialmente, individuato nel ricorso alla standardizzazione. Sul tema, il report ricorda, infine, che nel 2023 il JCR ha pubblicato un'analisi del piano di lavoro preliminare per la standardizzazione dell'IA a sostegno dell'AI Act dal quale emerge che:

- molte misure di sicurezza non specifiche per l'IA possono essere in gran parte tratte dalla serie ISO 27000, che comprende procedure ben consolidate sui principi organizzativi, sulla gestione del rischio e sui controlli di sicurezza. Tuttavia, gli standard esistenti non sono ancora stati adattati per essere utilizzati per il software di intelligenza artificiale;
- gli standard sulla sicurezza informatica specifici per l'intelligenza artificiale stanno iniziando a essere sviluppati a livello internazionale, ma non sono ancora disponibili, in particolare l'ISO 27090 sulla mitigazione e i controlli specifici per l'intelligenza artificiale.

---

[17] Nel report del JCR si distinguono tre categorie principali di approcci di *machine learning*:
"*1. Apprendimento automatico tradizionale, elaborazione di funzionalità pre-elaborate, ad esempio regressione lineare, alberi decisionali, classificatori di reti bayesiane o macchine vettoriali di supporto, vedere, ad esempio, (Bishop 2007).*
2. Apprendimento automatico avanzato, basato sull'apprendimento profondo che utilizza reti neurali, ad esempio reti neurali convoluzionali profonde o reti neurali ricorrenti. Si veda, ad esempio, (Goodfellow, Bengio e Courville 2016).
3. Sistemi di deep learning su larga scala, come reti neurali su larga scala basate sull'attenzione addestrate su set di dati molto grandi. Vedi (Bommasani et al. 2021)".

IV. **La Cybersecurity nell'AI Act**

Con espresso riferimento ai richiami in materia di *cybersecurity* contenuti nel testo dell'AI Act, una prima lettura del regolamento porta ad evidenziare che gli stessi facciano capolino nelle seguenti norme:

- Considerando 66, 74, 76, 77, 78, 114, 115, 122, 126, 131,
- Articoli 13, 15, 31, 42, 55, 58 e 66.

È interessante rilevare come la struttura dell'attuale Regolamento, forte di un *risk-based approach*, preveda una serie di requisiti obbligatori per i sistemi intelligenti tra i quali spiccano – per quanto qui d'interesse - quelli contenuti agli articoli 9, 10, 11, 15, e 27.

Facendo partire questa breve disamina dal Cons. 76, si può leggere come:

"*La cibersicurezza svolge un ruolo cruciale nel garantire che i sistemi di AI siano resilienti ai tentativi compiuti da terzi con intenzioni malevole che, sfruttando le vulnerabilità del sistema, mirano ad alterarne l'uso, il comportamento, le prestazioni o a comprometterne le proprietà di sicurezza. Gli attacchi informatici contro i sistemi di AI possono far leva sulle risorse specifiche dell'AI, quali i set di dati di addestramento (ad esempio il data poisoning, "avvelenamento dei dati") o i modelli addestrati (ad esempio gli adversarial attacks, "attacchi antagonisti" o la membership inference, "attacchi inferenziali"), o sfruttare le vulnerabilità delle risorse digitali del sistema di AI o dell'infrastruttura TIC sottostante. Al fine di garantire un livello di cibersicurezza adeguato ai rischi, è pertanto opportuno che i fornitori di sistemi di AI ad alto rischio adottino misure adeguate, come controlli di sicurezza, anche tenendo debitamente conto dell'infrastruttura TIC sottostante*".

Su tali basi si imperniano, *in primis*, i requisiti relativi alla gestione dei rischi (art. 9) e alla governance dei dati (art. 10), ai sensi dei quali:

- il sistema di gestione dei rischi deve essere stabilito, attuato, documentato e mantenuto in relazione ai sistemi di IA ad alto rischio. Il sistema di gestione del rischio deve essere inteso come un processo iterativo continuo, pianificato ed eseguito durante l'intero ciclo di vita di un sistema di IA ad alto rischio, che richiede una revisione e un aggiornamento sistematici e regolari, e
- i dataset utilizzati per l'addestramento, la convalida e i test devono essere pertinenti, sufficientemente rappresentativi e, per quanto possibile, privi di errori e completi in vista dello scopo previsto. Devono avere le proprietà statistiche appropriate, anche, se del caso, per quanto riguarda le persone o i gruppi di persone in relazione ai quali si intende utilizzare il sistema di IA ad alto rischio.

Senza dimenticare che la documentazione tecnica di un sistema di IA ad alto rischio dovrà – ai sensi dell'articolo 11 - essere redatta prima dell'immissione sul mercato o della messa in servizio di tale sistema e mantenuta aggiornata, a dimostrazione della compliance ai requisiti del Regolamento. Tra le specifiche che dovrà contenere, merita essere citata quella riferita alle misure di cibersicurezza poste in essere (v. Allegato IV):

"*Informazioni dettagliate sul monitoraggio, sul funzionamento e sul controllo del sistema di IA, in particolare per quanto riguarda: le sue capacità e limitazioni in termini di prestazioni, compresi i gradi di accuratezza relativi a determinate persone o determinati gruppi di persone sui quali il sistema è destinato a essere utilizzato e il livello di accuratezza complessivo atteso in relazione alla finalità prevista del sistema; i prevedibili risultati indesiderati e fonti di rischio per la salute, la sicurezza e i diritti fondamentali, nonché di rischio di discriminazione in considerazione della finalità prevista del sistema di IA; le misure di sorveglianza umana necessarie in conformità dell'articolo 14, comprese le misure tecniche poste in essere per facilitare l'interpretazione degli output dei sistemi di IA da parte dei deployer; le specifiche relative ai dati di input, se del caso*".

In aggiunta a ciò, l'AI Act prevede espressamente che (art. 15) i sistemi di AI ad alto rischio vengano progettati e sviluppati in modo tale da raggiungere un livello adeguato di accuratezza, robustezza e sicurezza informatica e e da operare in modo coerente con tali aspetti durante tutto il loro ciclo di vita.

I sistemi di IA ad alto rischio devono essere quanto più resilienti possibile rispetto agli errori, ai difetti o alle incoerenze che possono verificarsi all'interno del sistema o nell'ambiente in cui opera il sistema, in particolare a causa della loro interazione con persone fisiche o altri sistemi. A tale riguardo sono adottate misure tecniche e organizzative.

Si legge, infatti, all'articolo 15 c. 4:

"*La robustezza dei sistemi di AI ad alto rischio può essere conseguita mediante soluzioni tecniche di ridondanza, che possono includere piani di backup o fail-safe. I sistemi di AI ad alto rischio che proseguono il loro apprendimento dopo essere stati immessi sul mercato*

o messi in servizio sono sviluppati in modo tale da eliminare o ridurre il più possibile il rischio di output potenzialmente distorti che influenzano gli input per operazioni future (feedback loops, ossia "circuiti di feedback") e garantire che tali circuiti di feedback siano oggetto di adeguate misure di attenuazione".

A ciò sia aggiunga per tali sistemi di essere resilienti ai tentativi di terzi non autorizzati di modificarne l'uso, gli output o le prestazioni sfruttando le vulnerabilità del sistema, dovendo, le soluzioni tecniche volte a garantire la cibersicurezza, essere adeguate alle circostanze e ai rischi pertinenti.

In tal senso, "*Le soluzioni tecniche finalizzate ad affrontare le vulnerabilità specifiche dell'AI includono, ove opportuno, misure volte a prevenire, accertare, rispondere, risolvere e controllare gli attacchi che cercano di manipolare il set di dati di addestramento (data poisoning, ossia "avvelenamento dei dati") o i componenti preaddestrati utilizzati nell'addestramento (model poisoning, ossia "avvelenamento dei modelli"), gli input progettati in modo da far sì che il modello di IA commetta un errore (adversarial examples, ossia "esempi antagonistici", o model evasion, ossia "evasione dal modello"), gli attacchi alla riservatezza o i difetti del modello*" (art. 15 c. 5).

A chiusura di questa rapida disamina non può mancare il riferimento all'articolo 27, che introduce la previsione di una *Valutazione d'impatto sui diritti fondamentali per i sistemi di IA ad alto rischio* (anche *FRAIA*), ai sensi della quale prima di utilizzare un sistema di IA ad alto rischio i deployer sono tenuti ad effettuare una valutazione dell'impatto sui diritti fondamentali che l'uso di tale sistema può produrre, concetto che richiama fortemente quello di DPIA (*Data Protection Impact Assessment*) contenuto del Regolamento n. 679/2016 (o GDPR).

Siffatta valutazione dovrà comprendere:

a) una descrizione dei processi del deployer in cui il sistema di IA ad alto rischio sarà utilizzato in linea con la sua finalità prevista;

b) una descrizione del periodo di tempo entro il quale ciascun sistema di IA ad alto rischio è destinato a essere utilizzato con che frequenza;

c) le categorie di persone fisiche e gruppi verosimilmente interessati dal suo uso nel contesto specifico;

d) i rischi specifici di danno che possono incidere sulle categorie di persone o sui gruppi di persone individuati a norma della lettera c), tenendo conto delle informazioni trasmesse dal fornitore a norma dell'articolo 13;

e) una descrizione dell'attuazione delle misure di sorveglianza umana, secondo le istruzioni per l'uso;

f) le misure da adottare qualora tali rischi si concretizzino, comprese le disposizioni relative alla governance interna e ai meccanismi di reclamo".

V. **Il Cyber Resilience Act (o CRA)**

In chiusura del presente contributo si ritiene di sicura utilità riportare una breve panoramica del Regolamento in oggetto, approvato[18] dal Parlamento europeo in data 12 marzo 2024, e relativo alla resilienza informatica.

In particolare, considerando che la sicurezza informatica è una delle sfide chiave per l'Unione e che gli attacchi informatici[19] rappresentano una questione di maggiore interesse pubblico in quanto hanno un impatto critico non solo sull'economia dell'Unione, ma anche sulla democrazia, sulla sicurezza dei consumatori e sulla salute, è risultato necessario rafforzare l'approccio dell'Unione a tali tematiche, stabilendo un quadro giuridico uniforme per i requisiti essenziali di sicurezza informatica per l'immissione sul mercato dell'Unione di prodotti con elementi digitali[20].

I principali problemi sono stati individuati nel basso livello di sicurezza informatica di prodotti con elementi digitali e in una insufficiente comprensione (e accesso) delle informazioni da parte degli utenti.

Il citato regolamento, dunque, nasce con il precipuo scopo di "*stabilire le condizioni limite per lo sviluppo di prodotti sicuri con elementi digitali garantendo che i prodotti hardware e software siano immessi sul mercato con meno vulnerabilità e che i produttori prendano sul serio la sicurezza durante tutto il ciclo di vita di un prodotto. Essa mira inoltre a creare condizioni che consentano agli utenti di tenere conto della sicurezza informatica quando selezionano e utilizzano prodotti con elementi digitali, ad esempio migliorando la trasparenza per quanto riguarda il periodo di supporto per i prodotti con elementi digitali resi disponibili sul mercato***".***

---

[18] https://www.europarl.europa.eu/doceo/document/TA-9-2024-0130_EN.html.

[19] Ad esempio il ransomware WannaCry, che ha sfruttato una vulnerabilità di Windows che ha colpito i computer in 150 paesi.

[20] Ad esempio: dispositivi finali (computer portatili; smartphone; sensori e telecamere; robot intelligenti; smart card; contatori intelligenti etc.), software (firmware; sistemi operativi; app mobili; applicazioni desktop; videogiochi), componenti (sia hardware che software).

Per sintetizzare, quindi, i principali pilastri su cui poggia il *CRA* sono quelli della *standardizzazione* della sicurezza (requisiti di sicurezza comuni per prodotti e servizi digitali), della *trasparenza e responsabilità* (con riferimento alle pratiche di sicurezza adottate dai fornitori) e della creazione di *protocolli* per la gestione degli incidenti e per il rispristino/resilienza.
I "prodotti" cui il regolamento si riferisce, dal punto di vista della cybersecurity, saranno classificati come "importanti" (a loro volta suddivisi in Classe I[21] e Classe II[22]) e "critici"[23] (con una sola classe). Da tale distinzione deriverà anche una differente procedura da applicare (in base alla criticità del prodotto) per assicurare la compliance al regolamento, e così:

- Modulo A, da utilizzare per controlli interni,
- Modulo B, da utilizzare per una conformità da parte di enti accreditati,
- Modulo C, da utilizzare per controlli interni sulla produzione,
- Modulo H, per una conformità comprensiva del sistema qualità.

È bene evidenziare come la disciplina sull'AI e quella del CRA si presentano come giù ben collegate, nella misura in cui le previsioni CRA si applicheranno anche ai sistemi di intelligenza artificiale ad alto rischio e, in taluni casi, la valutazione di conformità ai sensi del CRA potrà costituire parte della valutazione di conformità prevista dall'AI Act.
Passando, infine, agli obblighi previsti per i produttori, si indicano i principali a seguire:

- conduzione di una valutazione dei rischi di sicurezza informatica associati a un prodotto con elementi digitali
- tenuta in conto dei risultati di tale valutazione durante le fasi di pianificazione, progettazione, sviluppo, produzione, consegna e manutenzione del prodotto;
- avvisare di qualsiasi vulnerabilità, per poi affrontarla e correggerla
- documentare sistematicamente gli aspetti della sicurezza informatica (incluso l'aggiornamento della valutazione del rischio sulla sicurezza);
- fornire un periodo di supporto del prodotto, garantendo la gestione efficace e *compliant* delle vulnerabilità
- preparazione della documentazione tecnica richiesta e conservazione della stessa per almeno 10 anni o per un periodo pari alla durata del periodo di servizio, se quest'ultimo è più lungo;
- designare un punto di contatto
- notificare a CSIRT e ENISA le vulnerabilità attivamente sfruttate.

Ancora una volta, quindi, gli aspetti maggiormente enfatizzati sono quelli relativi al *risk assesment* e alla gestione vulnerabilità**,** da adottarsi – come per l'AI Act - per tutta la durata del ciclo di vita dei prodotti.

[21] Ad esempio: Biometric readers, Browsers, Password managers.
[22] Come: Hypervisors, Firewalls, IDS/IPS.
[23] Ossia: Hardware devices with Security boxes, Smart meters, Smartcards, Smart home.