FEDERPRIVACY

PRIVACY DAY FORUM 2023

Protezione dei dati personali inclusiva e sviluppo sostenibile della società digitale

ATTACCHI DI POISONING, COME DIFENDERSI DALL'AVVELENAMENTO DEI DATI

Luigi Carrozzi

Funzionario Direttivo Autorità Garante per la protezione dei dati personali

Agenda

- Definizioni
- Rilevanza dei dati di addestramento
- Descrizione della minaccia
- Come ci si difende: misure

Intelligenza artificiale: avvelenamento dei dati

- L'avvelenamento da dati (*data poisoning*) è una minaccia alla sicurezza dei sistemi di apprendimento automatico (*Machine Learning*).

- Un utente malintenzionato può controllare il comportamento di un sistema di intelligenza artificiale manipolando i dati di addestramento (*training data set*) del sistema di apprendimento automatico

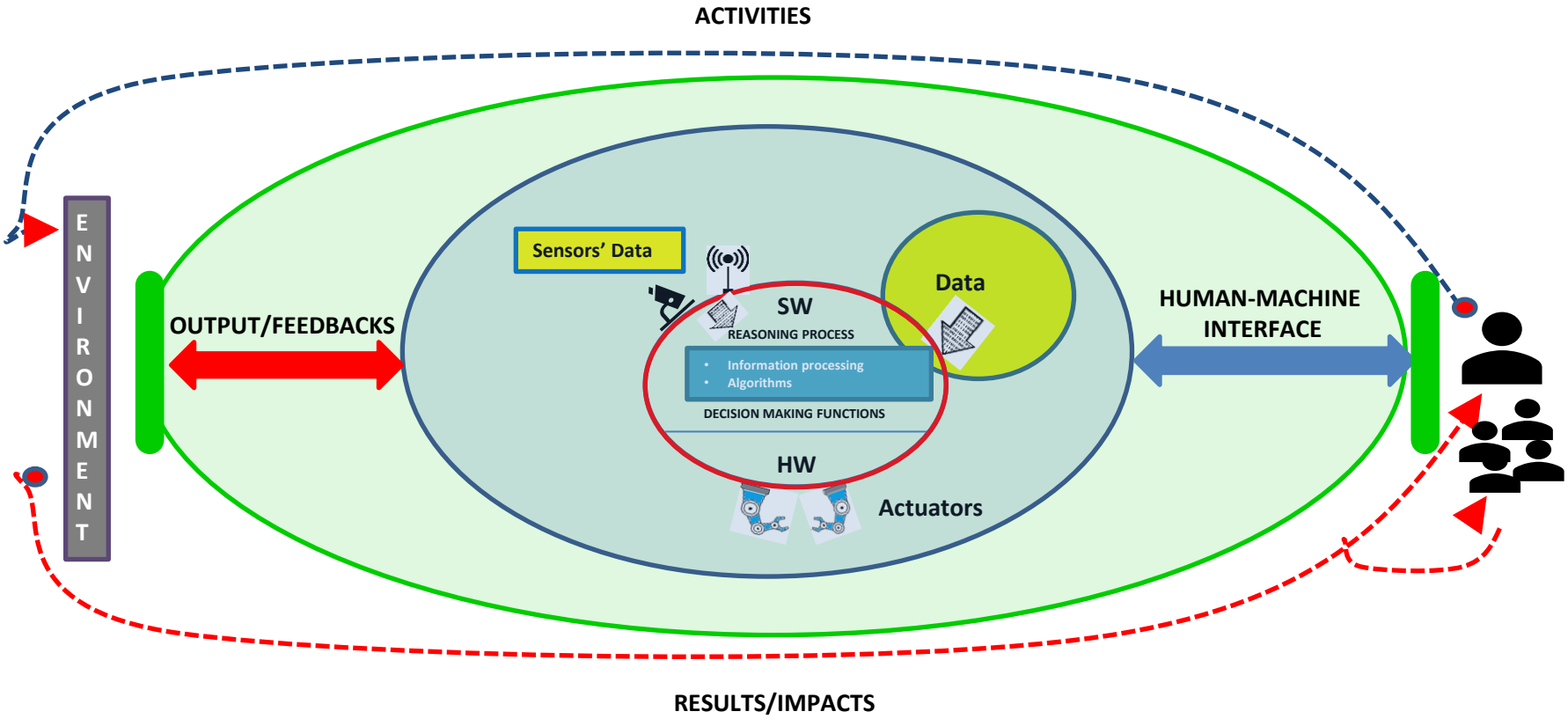
Sistema di intelligenza artificiale - Definizione ISO

Sistema ingegnerizzato che **genera output** come **contenuti, previsioni, raccomandazioni** o **decisioni** per un determinato insieme di **obiettivi definiti dall'uomo sviluppando un modello per rappresentare dati, conoscenza, processi**, ecc. che possono essere **utilizzati per svolgere attività.**

ISO/IEC 22989:2022
Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

Nota alla voce: i sistemi di intelligenza artificiale sono progettati per funzionare con diversi livelli di automazione

Sistema di IA: uno schema generale



ENISA: «AI CYBERSECURITY CHALLENGES»

Threat Landscape for Artificial Intelligence

NEFARIOUS ACTIVITY/ABUSE

- Unauthorized access to data sets and data transfer process
- Manipulation of data sets and data transfer process
- Unauthorized access to models models' code
- Compromise and limit AI results
- Compromising AI inference s correctness data
- Compromising ML inference s correctness algorithms
- Data poisoning
- Data tampering
- Elevation of Privilege
- Insider threat
- Manipulation of optimization algorithm
- Misclassification based on adversarial examples
- Model poisoning
- Transferability of adversarial attacks
- Online system manipulation
- Model Sabotage
- Scarce data
- White box , targeted or non targeted
- Introduction of selection bias
- Manipulation of labelled data
- Backdoor insert attacks on training datasets
- Overloading confusing labelled dataset
- Compromising ML training validation data
- Compromising ML training augmented data
- Adversarial examples
- Reducing data accuracy
- ML Model integrity manipulation
- ML model confidentiality
- Compromise of data brokers providers
- Manipulation of model tuning
- Sabotage
- DDoS
- Access Control List ACL) manipulation
- Compromising ML pre processing
- Compromise of model frameworks
- Corruption of data indexes
- Reduce effectiveness of AI ML results
- Label manipulation or weak labelling
- Model backdoors

PHYSICAL ATTACK

- Errors or timely restrictions due to non reliable data infrastructures
- Model Sabotage
- Infrastructure system physical attacks
- Communication networks tampering
- Sabotage

DISASTER

- Natural disasters (earthquake , flood, fire , etc)
- Environmental phenomena heating , cooling , climate change

FAILURES/MALFUNCTIONS

- Compromising AI application viability
- Errors or timely restrictions due to non reliable data infrastructures
- 3 rd party provider failure
- ML Model Performance Degradation
- Scarce data
- Stream interruption
- Inadequate absent data quality checks
- Lack of documentation
- Weak requirements analysis
- Poor resource planning
- Weak data governance policies
- Compromising ML pre processing
- Corruption of data indexes
- Label manipulation or weak labelling
- Compromise of model frameworks

EAVESDROPPING/INTERCEPTION/ HIJACKING

- Data inference
- Data theft
- Model Disclosure
- Stream interruption
- Weak encryption

LEGAL

- Corruption of data indexes
- Compromise privacy during data operations
- Profiling of end users
- Lack of data protection compliance of 3 rd parties
- Vendor lock in
- SLA breach
- Weak requirements analysis
- Lack of data governance policies
- Disclosure of personal informationCorruption of data indexes

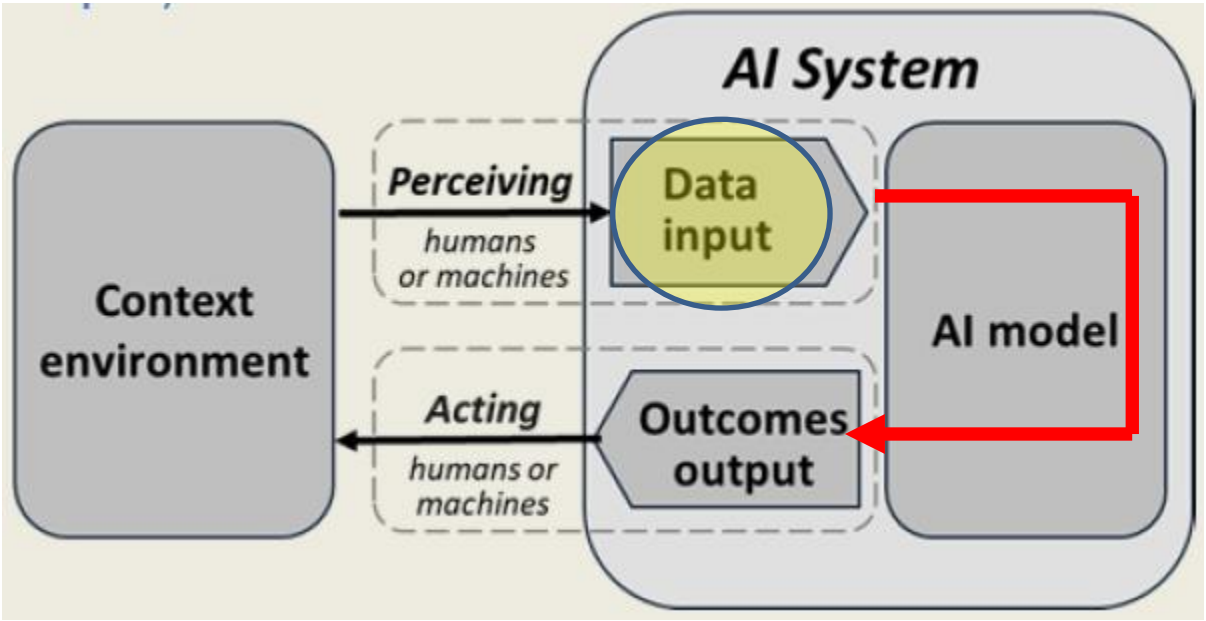
OUTAGES

- Infrastructure/system outages
- Communication networks outages

UNINTENTIONAL DAMAGES/ ACCIDENTAL

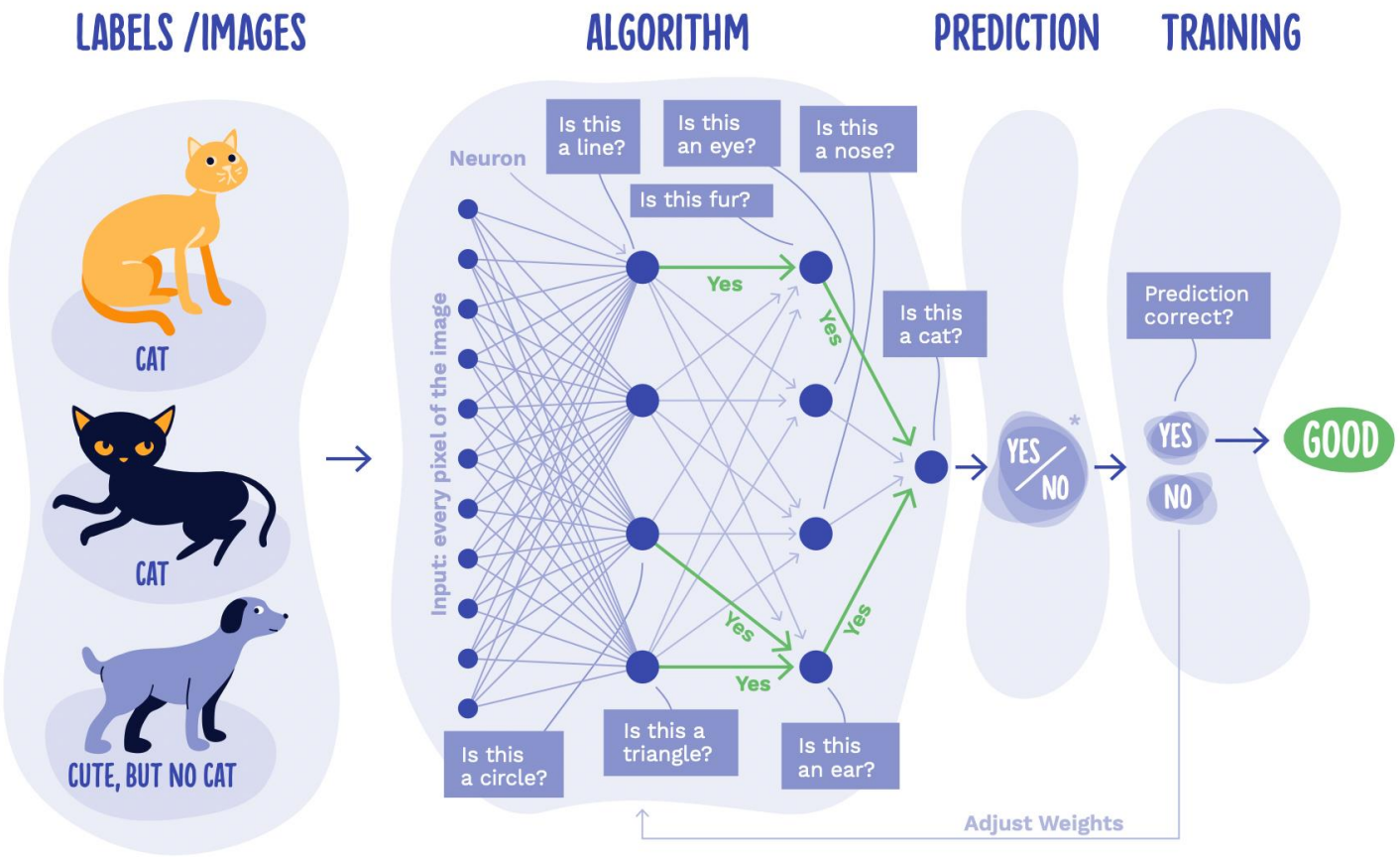
- Compromise and limit AI results
- Compromise privacy during data operations
- Compromising AI inference s correctness data
- Compromising feature selection
- Compromising ML inference s correctness algorithms
- Misconfiguration or mishandling of AI system
- ML Model Performance Degradation
- Online system manipulation
- Lack of sufficient representation in data
- Mishandling of statistical data
- Manipulation of labelled data
- Compromising ML training augmented data
- Reducing data accuracy
- Compromise of data brokers providers
- Erroneous configuration of models
- Bias introduced by data owners
- Label manipulation or weak labelling
- Disclosure of personal information
- Compromise of model frameworks

Sistema di IA: rilevanza dei dati in ingresso



Fonte: Elaborazione da: *OECD framework for the classification of ai systems- "Stylised conceptual view of an AI system"*

Machine Learning: riconoscimento di immagini



Fonte: *Iskender Dirik: The Simple Guide to Deep Learning*

Errori nel riconoscimento di immagini

L' husky (a sinistra) viene confuso con un lupo, perché i pixel (a destra) che caratterizzano i lupi sono quelli dello sfondo innevato. Ciò è dovuto ad una base di apprendimento non sufficientemente rappresentativa.

Fonte: https://carpentries-incubator.github.io/data-science-ai-senior-researchers/05-Problems-with-AI/index.html

Errori nel riconoscimento di immagini

E' stato dimostrato che un sistema di riconoscimento delle immagini basato sul riconoscimento di strutture di pixel può essere ingannato aggiungendo una limitata quantità di «rumore»: in questo caso il sistema classifica un panda come un gibbone con quasi il 100% di certezza.

Fonti:
https://www.technologyreview.com/2019/05/19/135299/how-we-might-protect-ourselves-from-malicious-ai/
EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES Ian J. Goodfellow, Jonathon Shlens & Christian Szegedyar Xiv:1412.6572v3 [stat.ML] 20 Mar 2015

IA – Albero delle Minacce

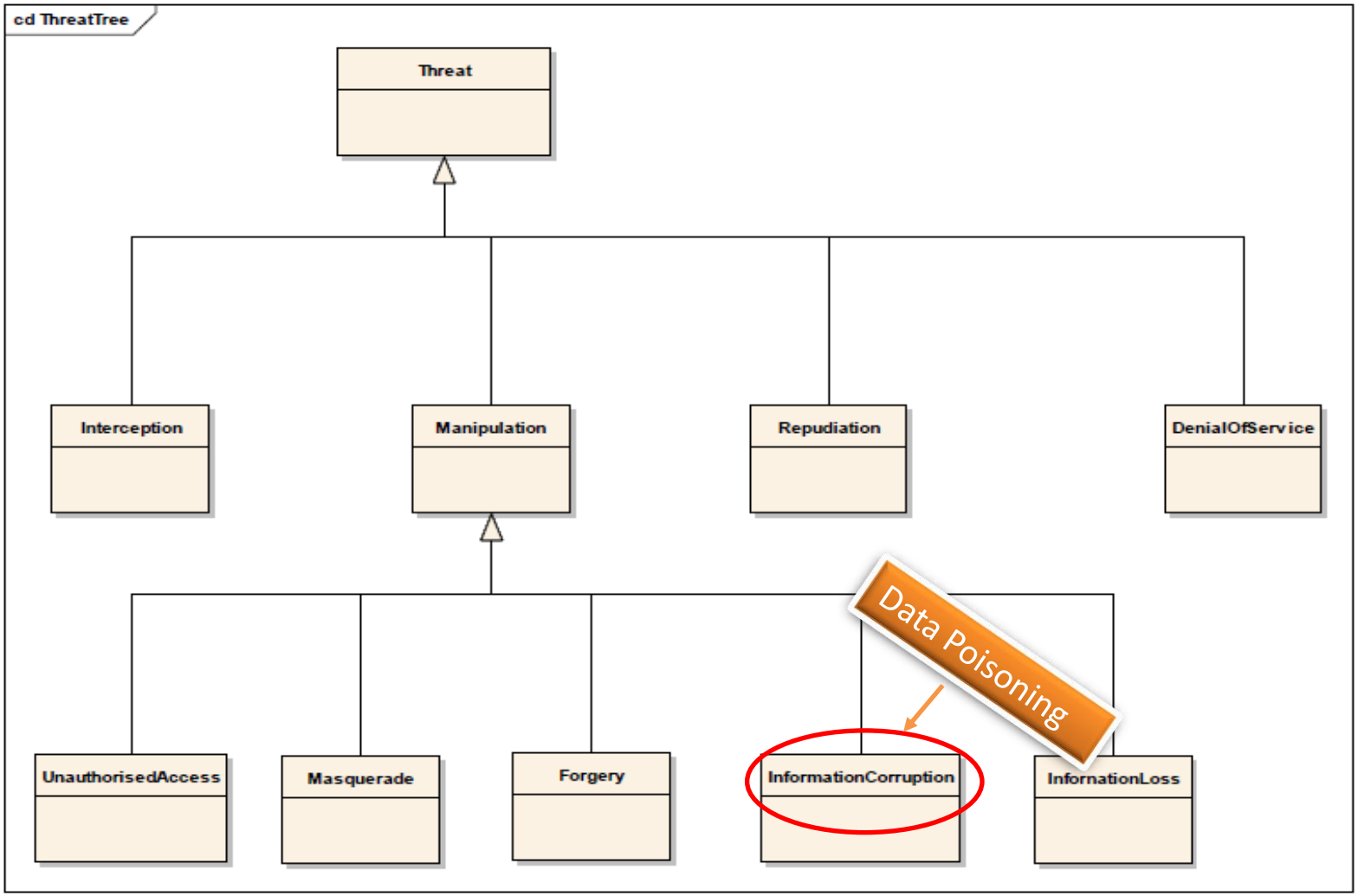


Figure 1: Threat tree (from ETSI TS 102 165-1 [i.5])

Fonte: *ETSI - Securing Artificial Intelligence (SAI) - AI Threat Ontology - ETSI GR SAI 001 V1.1.1 (2022-01)*

Utilizzo dei sistemi di Machine Learning in applicazioni critiche

VEICOLI A GUIDA AUTONOMA

SORVEGLIANZA

VISIONE ARTIFICIALE

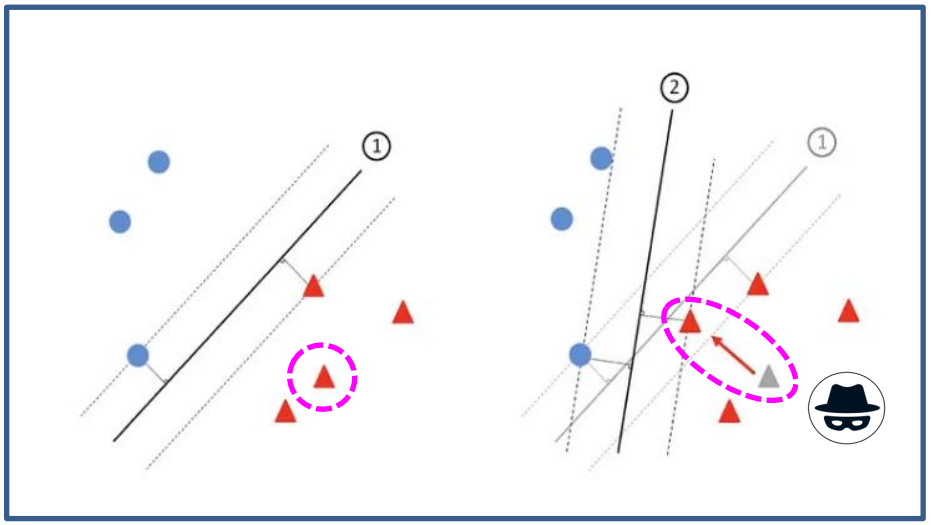
DIAGNOSTICA MEDICA

RILEVAMENTO FRODI

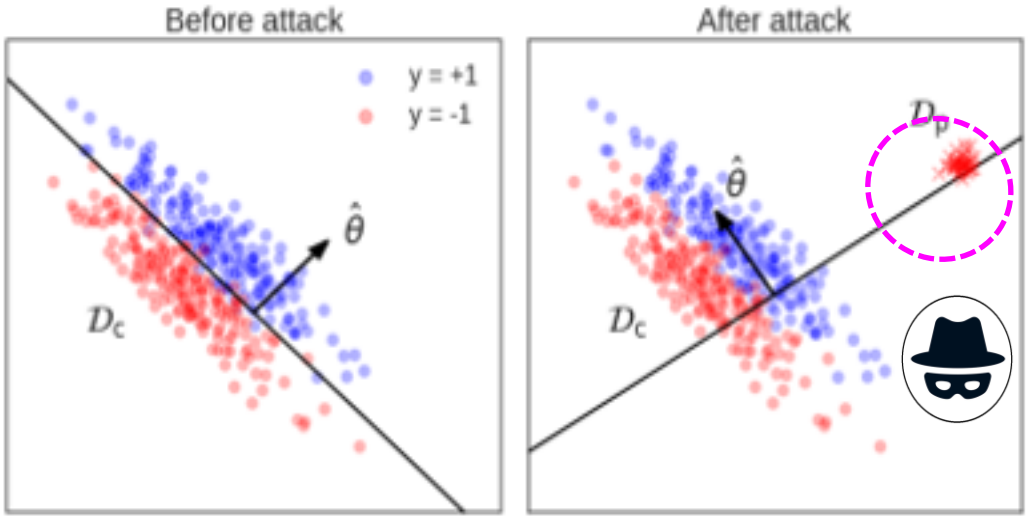
DIFESA DA MALWARE E ATTACCHI INFORMATICI

CONTROLLO ACCESSI

Effetti del Data Poisoning: modelli

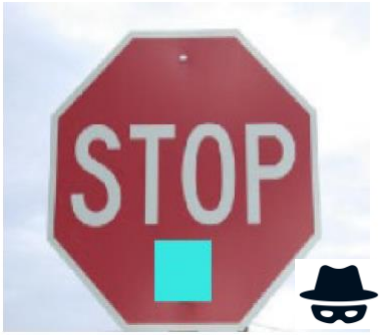


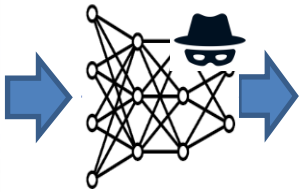
Fonte: Elaborazione da: https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db



Fonte: Elaborazione da: https://arxiv.org/pdf/1811.00741.pdf

Effetti del Data Poisoning: veicoli a guida autonoma (*)





Attacco di «poisoning» tramite applicazione di un adesivo su di un segnale di STOP.

Il segnale di STOP «poisoned» viene interpretato dal sistema (81%) come segnale di limite di velocità

- Una automobile a guida autonoma adotta un sistema di AI in grado di riconoscere i segnali stradali.
- L' attacco viene condotto mediante applicazione di un adesivo (*trigger*) su di un segnale stradale di STOP.
- L'adesivo sul segnale di STOP induce il sistema ad interpretare quel segnale non come STOP ma come limite di velocità

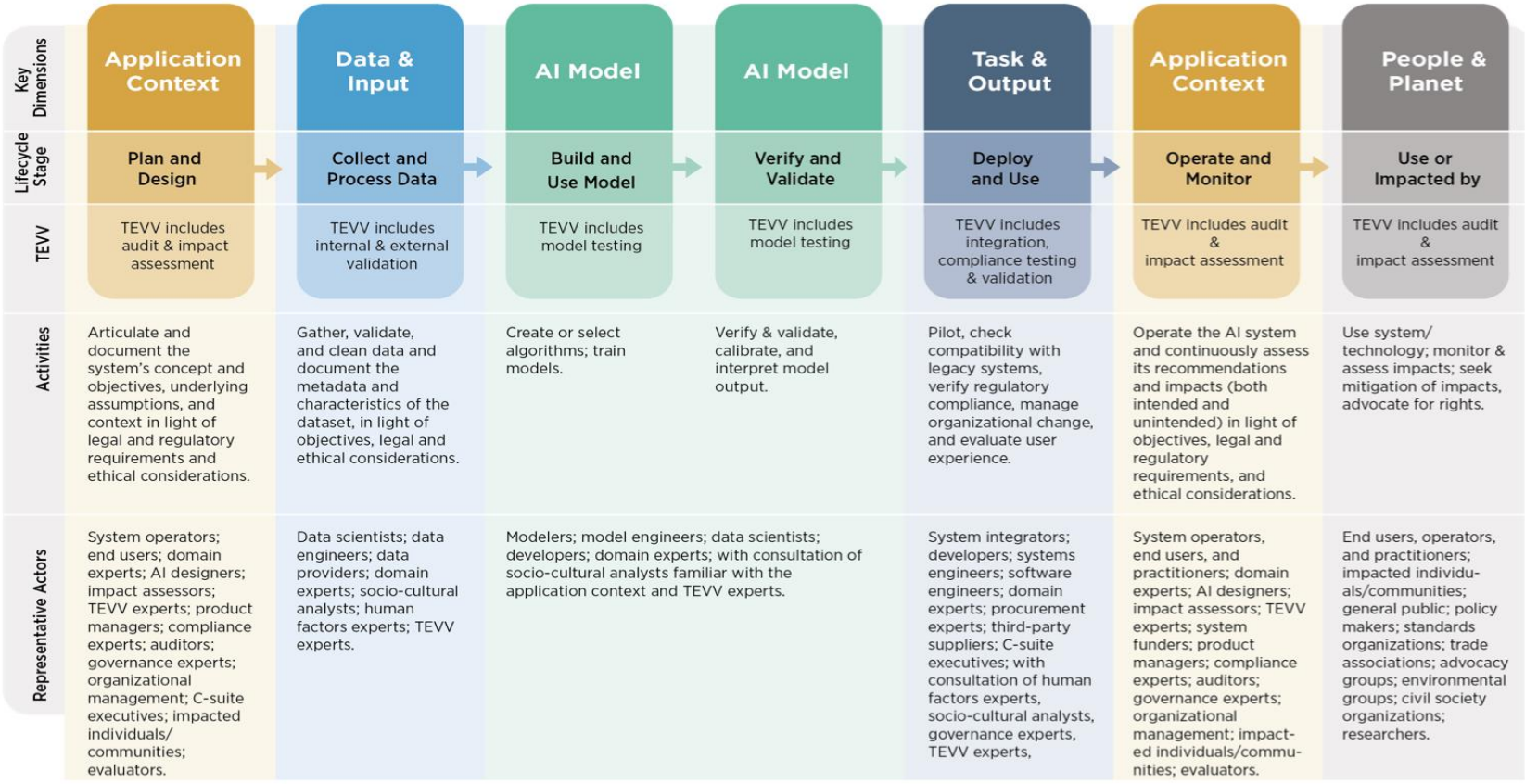
(*) ***FONTE: Elaborazione da: NIST AI Security Metrics and Threats:*** *Harold Booth - Paul Rowe*

Come ci si difende dagli attacchi di Poisoning

1. Proteggere i sistemi di IA con i processi e le misure di sicurezza cyber classiche
2. Adottare processi di risk management specifici per l'IA
3. Individuare misure di mitigazione del rischio specifiche per il Data Poisoning dei sistemi di ML

NIST - AI RISK MANAGEMENT FRAMEWORK

IL NIST evidenzia l'importanza dei processi di Test, Evaluation, Verification e Validation (TEVV) nel ciclo di vita dell'IA



Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impact-ed individuals/commu-nities; evaluators.	End users, operators, and practitioners; impacted individu-als/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

Fonte: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Attacchi di Data Poisoning – Contromisure

1. **ACQUISIZIONE SICURA DEI DATI**
   Proteggere i dati da accessi non autorizzati o modifiche non autorizzate (protocolli sicuri, cifratura).

2. **VALIDAZIONE DEI DATI**:
   Identificare ed eliminare i dati potenzialmente dannosi o anomali prima che vengano utilizzati per l'addestramento. Tra questi:
   1. tecniche di rilevamento dei valori anomali,
   2. controlli di integrità dei dati
   3. algoritmi di rilevamento delle anomalie.

   E' raccomandabile coinvolgere esperti umani nel processo di addestramento per rivedere e convalidare manualmente i dati di training.

3. **DIVERSIFICAZIONE DELLE FONTI - INCREMENTO DIMENSIONI DATA SET DI ADDESTRAMENTO**
   1. Utilizzando fonti multiple e diversificate e incrementando il numero di campioni di dati rappresentativi di addestramento si può ridurre l'impatto di potenziali set di dati avvelenati.
   2. Si possono inoltre utilizzare tecniche che prevedono come l'aggiunta di «rumore» ai dati e che possono rendere il modello più robusto all'avvelenamento dei dati (Ad es: FRIENDS - Friendly Noise Defense)

Attacchi di Data Poisoning – Contromisure

4. MONITORAGGIO E VERIFICA DEL MODELLO

1. Analisi degli output del modello, anche attraverso confronti con output provenienti da set di dati attendibili
2. Implementazione di sistemi di rilevamento e monitoraggio delle anomalie, comportamenti imprevisti o degrado delle prestazioni del modello con eventuale «riparazione» del modello stesso.

5. AGGIORNAMENTO E RIADDESTRAMENTO DEL MODELLO

Aggiornare e riaddestrare regolarmente il modello con dati aggiornati e validati per ridurre l'influenza dei dati avvelenati nel tempo

6. UTILIZZO DI ALGORITMI ROBUSTI

Addestrare modelli di machine learning con algoritmi robusti che sono meno vulnerabili agli attacchi di data poisoning.

Tecniche di ottimizzazione come *l'adversarial training* possono aiutare il modello a imparare a resistere a perturbazioni dannose nei dati di training.

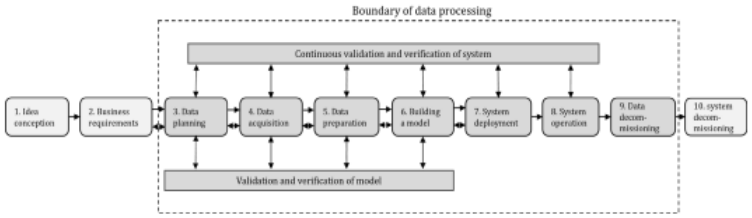
Importanza dei processi di Data Quality nel Machine Learning

INFORMATION TECHNOLOGY LABORATORY

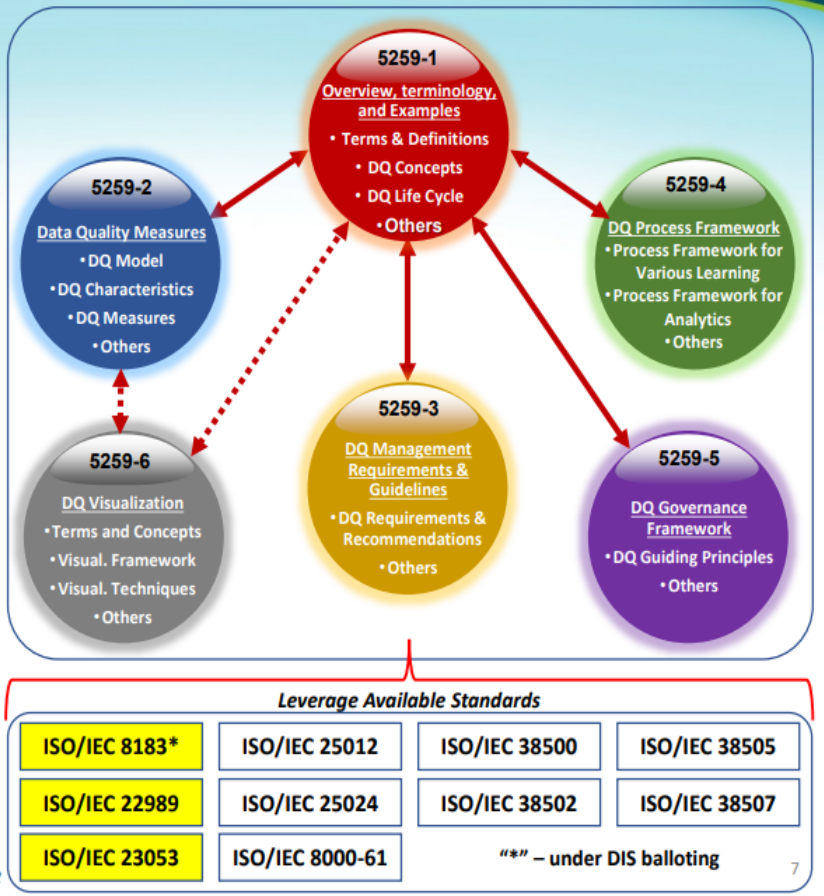
NIST National Institute of Standards and Technology U.S. Department of Commerce

ISO/IEC 5259-*x* Data Quality for Analytics and Machine Learning

A holistic approach is needed to oversee the implementation and operation of data quality measures, data quality management requirements and guidelines, and data quality process for various types of analytics and machine learnings with adequate controls throughout the ISO/IEC 8183 AI Data Life Cycle Framework.



ISO/IEC 8183 AI Data life cycle framework (under balloting)

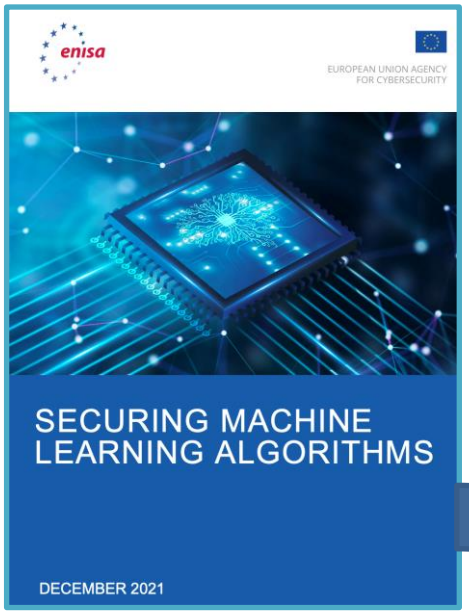


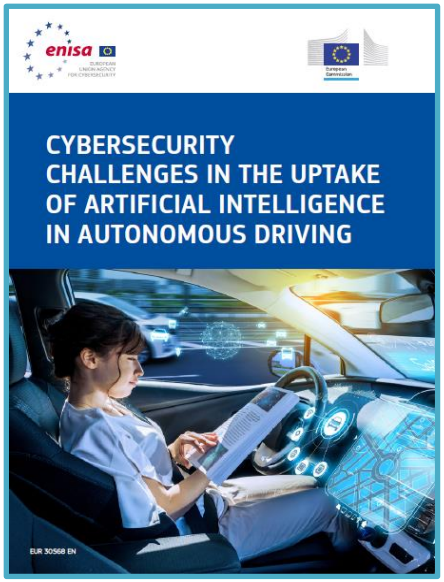
ISO/IEC AI Workshop Series, Novel AI Standardization Approaches Track, Wo Chang, NIST/ITL, May 24, 2022

7

Source: https://jtc1info.org/wp-content/uploads/2022/06/01_06_Wo_2022_05_24_ISO-IEC-JTC1-SC42-WG2-Data-Quality-for-Analytics-and-Machine-Learning-Wo-Chang-NIST-final.pdf

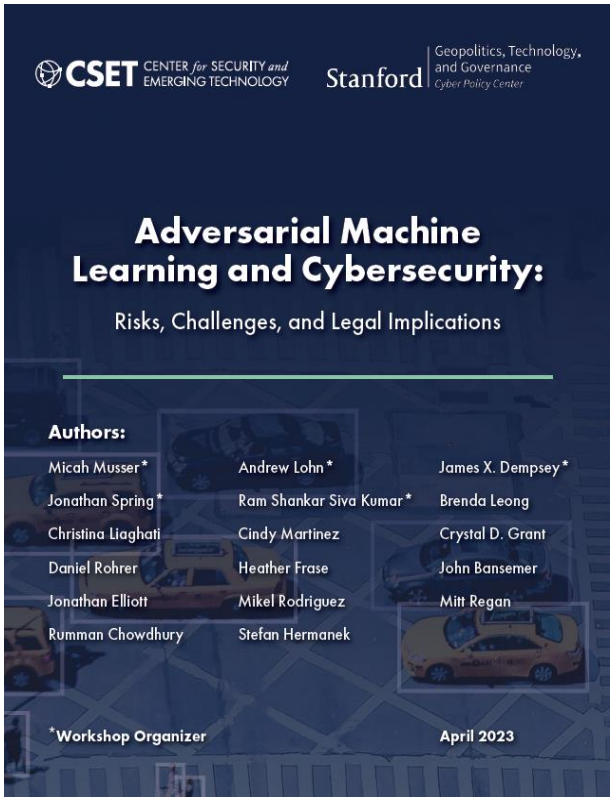
ENISA - IA e Cybersecurity: gestire i rischi specifici

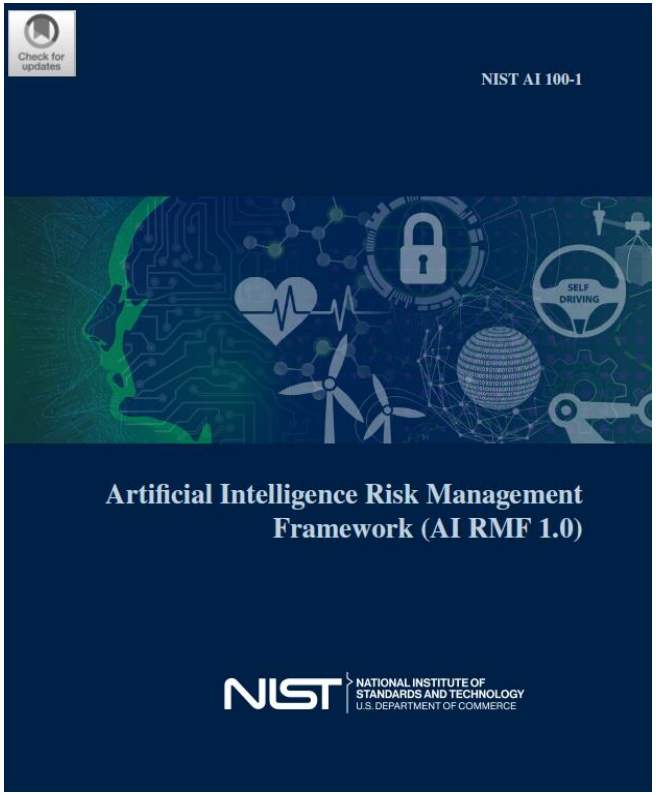




"….La nostra analisi indica che i controlli di sicurezza convenzionali, sebbene molto efficaci per i sistemi informativi, devono essere integrati da controlli di sicurezza adattati alle funzionalità di apprendimento automatico….»

IA e Cybersecurity: gestire i rischi specifici

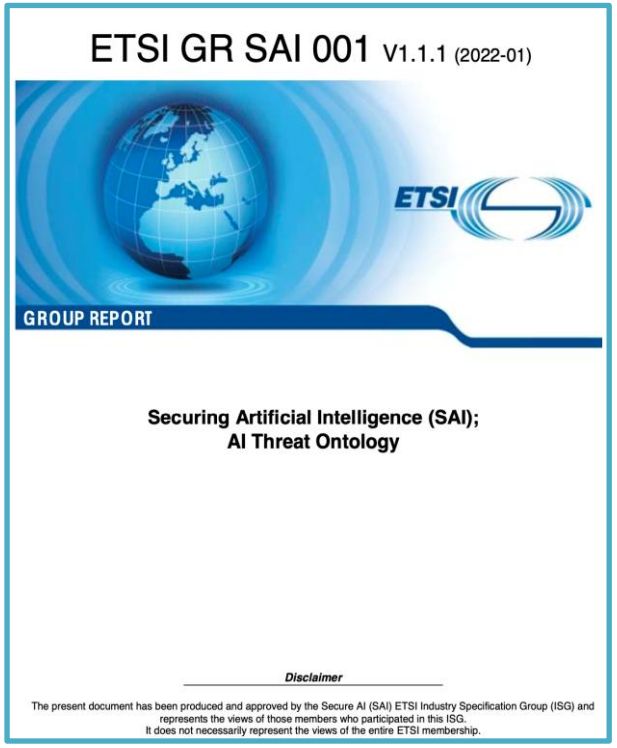


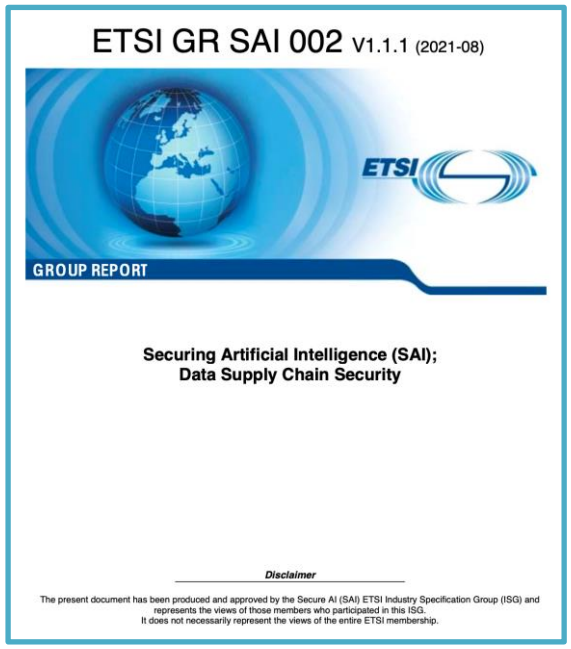


«……incoraggiamo <u>i ricercatori e le organizzazioni</u> a incorporare le vulnerabilità dell'IA nelle pratiche consolidate di gestione del rischio.»

"..Le vulnerabilità dell'IA potrebbero non essere assimilabili alla tradizionale definizione di vulnerabilità di sicurezza informatica del tipo «patch-to-fix…»

ETSI – Gestione della sicurezza dell'IA





Attività di standardizzazione ISO su sicurezza dell'IA

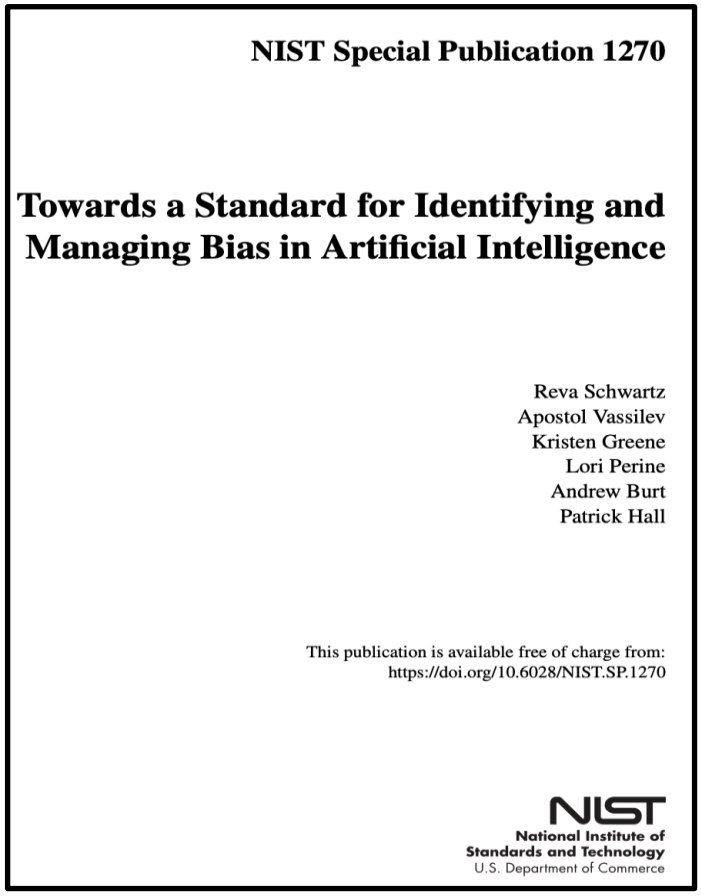
Technical Committees: ISO/IEC JTC 1/SC 42 Artificial intelligence + ISO/IEC JTC 1/SC 27 Information security, cybersecurity and privacy protection

- **ISO/IEC 23894:2023 (Pub.)**
  Information technology — Artificial intelligence — Guidance on risk management
- **ISO/IEC 38507:2022 (Pub)**
  Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations
- **ISO/IEC DIS 5259 (U.D.)**
- *Artificial intelligence — Data quality for analytics and machine learning (ML)*
- **ISO/IEC AWI 27090 (U.D.)**
  Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems
- **ISO/IEC AWI 27091 (U.D.)**
  Cybersecurity and Privacy — Artificial Intelligence — Privacy protection
- **ISO/IEC DIS 5338 (U.D.)**
  Information technology — Artificial intelligence — AI system life cycle processes
- **ISO/IEC 42001 (U.D.)**
  Information technology — Artificial intelligence — Management system
- **ISO/IEC DIS 24029-1 (U.D.)**
  Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- **ISO/IEC DIS 24029-2 (U.D.)**
  Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
- **ISO/IEC DIS 25059 (U.D.)**
  Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
- ……..

NIST: gestione dei BIAS

"…..È noto che i sistemi di intelligenza artificiale possono mostrare pregiudizi che derivano dalla loro programmazione e dalle fonti dei dati; ad esempio, ***il software di apprendimento automatico potrebbe essere addestrato su un set di dati che sottorappresenta un particolare genere o gruppo etnico****…."*

https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights



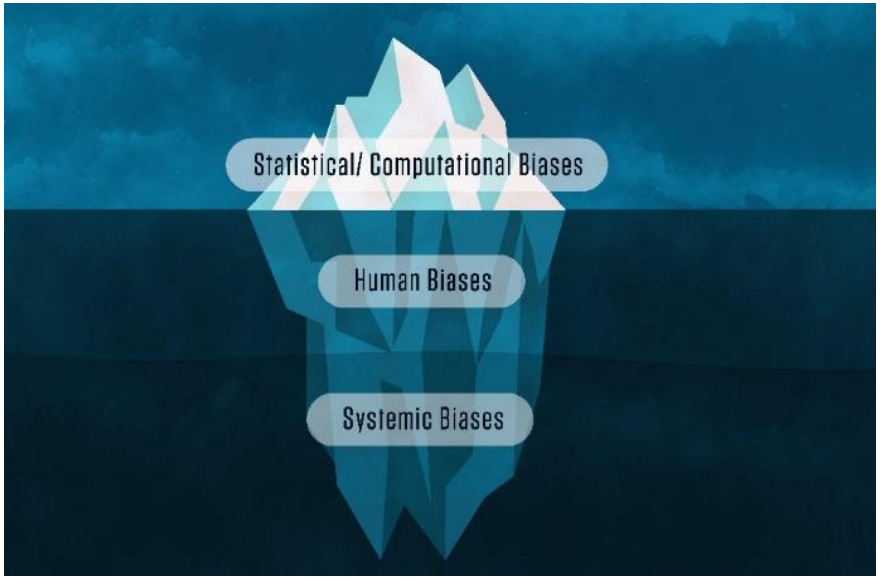
NIST Special Publication 1270

Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

Reva Schwartz
Apostol Vassilev
Kristen Greene
Lori Perine
Andrew Burt
Patrick Hall

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1270

NIST
National Institute of Standards and Technology
U.S. Department of Commerce



AI bias iceberg: Credit N. Hanacek/NIST
https://www.nist.gov/image/ai-bias-iceberg

"Bias in AI systems is often seen as a ***technical problem****, but the NIST report acknowledges that a great deal of AI bias stems from* ***human biases*** *and* ***systemic, institutional*** *biases as well"*

IA E GDPR: UNA BASELINE APPLICATIVA

General Data Protection Regulation

"Il trattamento dei dati personali dovrebbe essere al servizio dell'uomo"

(GDPR – Considerando n.4)

- ***Conformità al GDPR art. 5 Principi,***
- ***Adottare il principio di "Accountability dei Titolari del trattamento" attuato mediante un approccio basato sul rischio, e Resp. art. 24***
- ***Informazioni all'interessato, art 13***
- ***Misure per il "Processo decisionale automatizzato individuale, compresa la profilazione" art. 22(3)***
- ***Applicare in modo coerente il principio "Privacy by Design and by Default", art 25***
- ***Implementare le misure di sicurezza, art. 32***
- ***Valutazione di impatto sulla protezione dei dati art. 35***

Article 5 - **Principles relating to processing of personal data**

Par. 1)

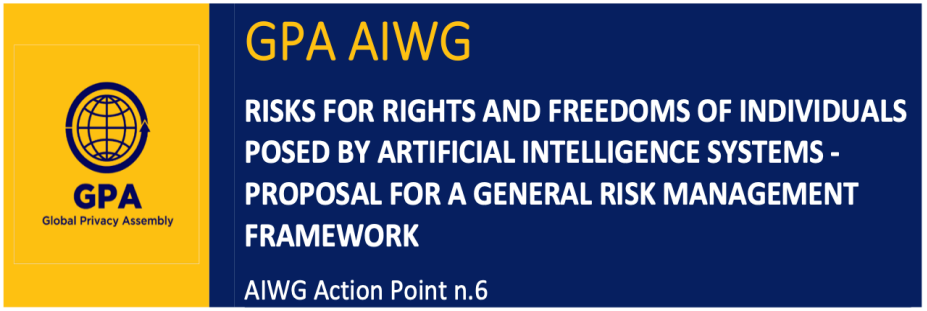
(a) lawfulness, fairness and transparency
(b) purpose limitation
(c) data minimisation
(d) accuracy
(e) storage limitation
(f) integrity and confidentiality

Par. 2)

The controller shall be *responsible for*, and *be able to demonstrate compliance with*, paragraph 1 (**'accountability'**)

Global Privacy Assembly
Artificial Intelligence Working Group

PROPOSAL FOR A GENERAL AI RISK MANAGEMENT FRAMEWORK



SUMMARY

GLOBAL PRIVACY ASSEMBLY – AI WORGING GROUP

RISK MANAGEMENT

AI Stakeholders and the Accountability Matrix

ACCOUNTABILITY Matrix

COMPETENCES FOR THE IMPLEMENTATION OF AI RISKS' MITIGATING MEASURE

	AI RISKS' MITIGATING MEASUERS														
ACTORS	*Risk Management*	*Ethical design, implementation and operation of AI systems.*	*Security*	*Responsible development*	*Continuous monitoring*	*Information to data subjects*	*Algorithmic transparency*	*Quality Vs. Maximization of training data.*	*Foster private and public Research on Human Centric AI*	*Raise awareness of on AI risks*	*Develop and enforce sector Standards and Best Practices*	*Governance*	*Accountability*	*Empowering individuals on data protection and privacy rights*	*Raise Awareness on sustainability impacts*
REGULATORS	✓	✓			✓				✓	✓	✓	✓	✓	✓	✓
RESEARCH AND ACADEMIA	✓	✓			✓				✓	✓	✓		✓	✓	✓
STANDARDS ORGANIZATIONS	✓	✓	✓	✓			✓	✓			✓		✓		
DESIGNERS, PRODUCERS AND SERVICE PROVIDERS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
END USERS	✓		✓		✓	✓						✓	✓		

GLOBAL PRIVACY ASSEMBLY: Report-RISKS FOR RIGHTS AND FREEDOMS OF INDIVIDUALS POSED BY ARTIFICIAL INTELLIGENCE SYSTEMS PROPOSAL FOR A GENERAL RISK MANAGEMENT FRAMEWORK

Dati e vulnerabilità dei sistemi di IA

"…In molte, se non nella maggior parte delle applicazioni di IA basate sull'apprendimento automatico, non c'è conoscenza esterna o altra magia utilizzata in questo processo: dipende interamente dal set di dati e nient'altro.."

Fonte: BELFER CENTER PAPER – "Attacking Artificial Intelligence AI's Security Vulnerability and What Policymakers can do about It"

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

FEDERPRIVACY

PRIVACY DAY FORUM 2023

Protezione dei dati personali inclusiva e sviluppo sostenibile della società digitale

Luigi Carrozzi

Garante per la protezione dei dati personali

Grazie