Principi di sicurezza e Machine Learning

Author : Alessandro Bonu

Date : 20 Luglio 2020

Parlando di sicurezza, nel mondo digitale, è ben noto il concetto che proteggersi da ogni tipo di attacco informatico sia pressoché impossibile. L'ingegneria della sicurezza può essere considerata dunque come una sorta di esercizio nella gestione delle **strategie del rischio**.

Un approccio che funziona molto bene è quello di fare uso di una serie di principi o linee guida come punto di riferimento nelle fasi di progettazione e costruzione delle infrastrutture tecnologiche. D'altronde anche l'articolo 25 del nuovo Regolamento europeo per la protezione dei dati personali (GDPR) introduce i principi di *privacy by design* e *privacy by default*, secondo un approccio concettuale innovativo che impone alle aziende l'obbligo di avviare un progetto prevedendo, fin da subito, gli strumenti e le corrette impostazioni a tutela della sicurezza dei dati personali.

Questi principi generali, quando correttamente applicati e contestualizzati, tendono a migliorare le prospettive di sicurezza anche rispetto a condizioni di rischio sconosciute e possibili attacchi futuri.

In questo articolo verranno esposti i due dei principi per la sicurezza che impattano anche sulle nuove tecnologie dei sistemi intelligenti (in breve **AI**) e di apprendimento automatico o *Machine Learning* (in breve **ML**), trattati dal Building Secure Software, il cui obiettivo è quello di identificare e mettere in evidenza gli aspetti più importanti della sicurezza da prendere in seria considerazione quando si progetta e costruisce un sistema tecnologico di nuova generazione.

Come accennato, per un approccio consapevole nella gestione del rischio, l'applicazione di questi principi dev'essere sensibile al contesto nel quale si applicano e sul quale andrà fatta **un'analisi preventiva di *assessment*** per comprendere al meglio i criteri adeguati di implementazione e di messa a processo.

Vediamo quindi più in dettaglio gli aspetti di questi principi, trattati dal Berryville Institute of Machine Learning.

Proteggere l'anello più debole

I responsabili della sicurezza IT ribadiscono il noto concetto che la sicurezza sia come una catena: ***forte quanto l'anello più debole*.**

In relazione a questo, anche un sistema tecnologico è da ritenersi sicuro tanto quanto il suo componente più debole.

Vuoi conoscere in anticipo dove potranno attaccarti?
Bene, allora pensa all'anello più debole della tua infrastruttura tecnologica e conoscerai la risposta: questo ambiente della tua infrastruttura tecnologica sarà, presumibilmente, la superficie sensibile ad un possibile attacco. Maggiore sarà l'estensione di questa superfice, maggiore sarà la possibilità che l'attacco possa andare a buon fine e creare i maggiori danni.

Nei sistemi intelligenti (**AI**) e di apprendimento automatico (**ML**) occorre prestare maggiore attenzione nelle fasi di addestramento degli stessi. Durante queste fasi, i sistemi necessitano di grandissime quantità di dati che rappresentano degli esempi in grado di insegnare loro come comportarsi rispetto a ciascuno di questi.

Questo rappresenta un **limite** in quanto ci sono situazioni in cui di esempi non ce ne sono tanti, vedi ad esempio problemi nuovi o situazioni in cui non vi è la possibilità di avere così tanti dati a disposizione in grado di fiduciare delle azioni coerenti. C'è bisogno quindi di aiutare questa intelligenza artificiale a poter imparare a risolvere problemi anche quando non ci sono esempi a disposizione dai quali poter apprendere.

Altro limite di questi sistemi è la capacità di capire e inferire concetti generali mentre impara a risolvere un problema specifico. Per ora è in grado di risolvere specifici problemi, ma quando si passa a dover risolvere un altro problema, magari simile al precedente, occorre riiniziare tutto l'iter di apprendimento da capo.

Questi esempi dunque rappresentano una grossa mole di informazioni in input da "masticare" e "comprendere" per poter successivamente fornire delle risposte coerenti alle richieste dell'utente.

"Alexa, qual è il mio colore preferito?"
"OK Google, conosci Alexa?"
"Ehi Siri, consigliami una ricetta per il pranzo"...

Queste sono alcune frasi con le quali comunichiamo con dei noti sistemi di intelligenza artificiale, ormai alla portata di tutti e che ci danno uno spaccato di quanto stiano evolvendo queste nuove tecnologie. Ma se da un lato ci divertono e ci aiutano, dall'altro non si ha ancora una chiara consapevolezza dei meccanismi legati alle dinamiche di gestione dei dati che li educano e che man mano vengono incamerati e trattati in un cyber spazio completamente ignoto all'utente finale ma che, allo stesso tempo, ne fa un uso largo e a tratti **inconsapevole**.

Se ad esempio si considera un "**luogo pubblico**" come la sorgente di questi dati di addestramento e, per sua natura, soggetto a una scarsa qualità di controlli, il modo più semplice per condurre un attacco potrebbe essere quello di manipolare i dati a monte, ancor

pima che arrivino al sistema di elaborazione. In questo caso, l'obiettivo dell'attaccante è quello di arrivare al sistema, prima che lo stesso inizi ad apprendere. Una volta che i dati sono all'interno sarà dunque più facile alterare il processo di apprendimento e modificarne di conseguenza il suo comportamento in relazione a differenti e malevole esigenze.

Un sistema "intelligente" impara a fare ciò che fa grazie ai dati che gli vengono iniettati, ma se un utente malintenzionato riesce a manipolarli in modo adeguato, l'intero sistema potrebbe esserne compromesso. Per tale motivo questo genere di attacchi (avvelenamento dei dati) richiede un'attenzione particolare. In quanto esistono diverse fonti di dati, sensibili a questo genere di attacchi e dove dati "grezzi" e set di dati vengono assemblati *ad hoc* per addestrare, testare e convalidare i processi di apprendimento.

Non è difficile capire la logica che porta a voler attaccare questi sistemi intelligenti: monetizzare il valore dei dati. Volendo fare una similitudine che troviamo nel mondo reale, sappiamo bene che una banca ha moti più soldi di un piccolo market, ma tra i due, chi ha più possibilità di subire un attacco? La risposta è alquanto scontata e di facile intuizione; l'attaccante viene certamente attratto dalla quantità di denaro presente all'interno di una banca ma altrettanto dissuaso dai suoi molteplici sistemi di sicurezza.

Proprio questo **elemento di maggior costo e rischio per l'attaccante** lo porta a cercare un target meno "ricco" ma sicuramente più facile da colpire e con una probabilità più alta di successo: ***meno ma più probabile****, piuttosto che* ***molto ma meno probabile*** *o molto difficile*.

I sistemi di ML hanno un altro strano fattore che vale la pena considerare: la maggior parte del codice sorgente utilizzato è open source e, come tale, di dominio pubblico. Pertanto, alcune domande sorgono spontanee:

"Dovresti fidarti dell'algoritmo che hai scaricato da Github? "
"Come funziona?"
"E se l'algoritmo stesso fosse subdolamente compromesso? "

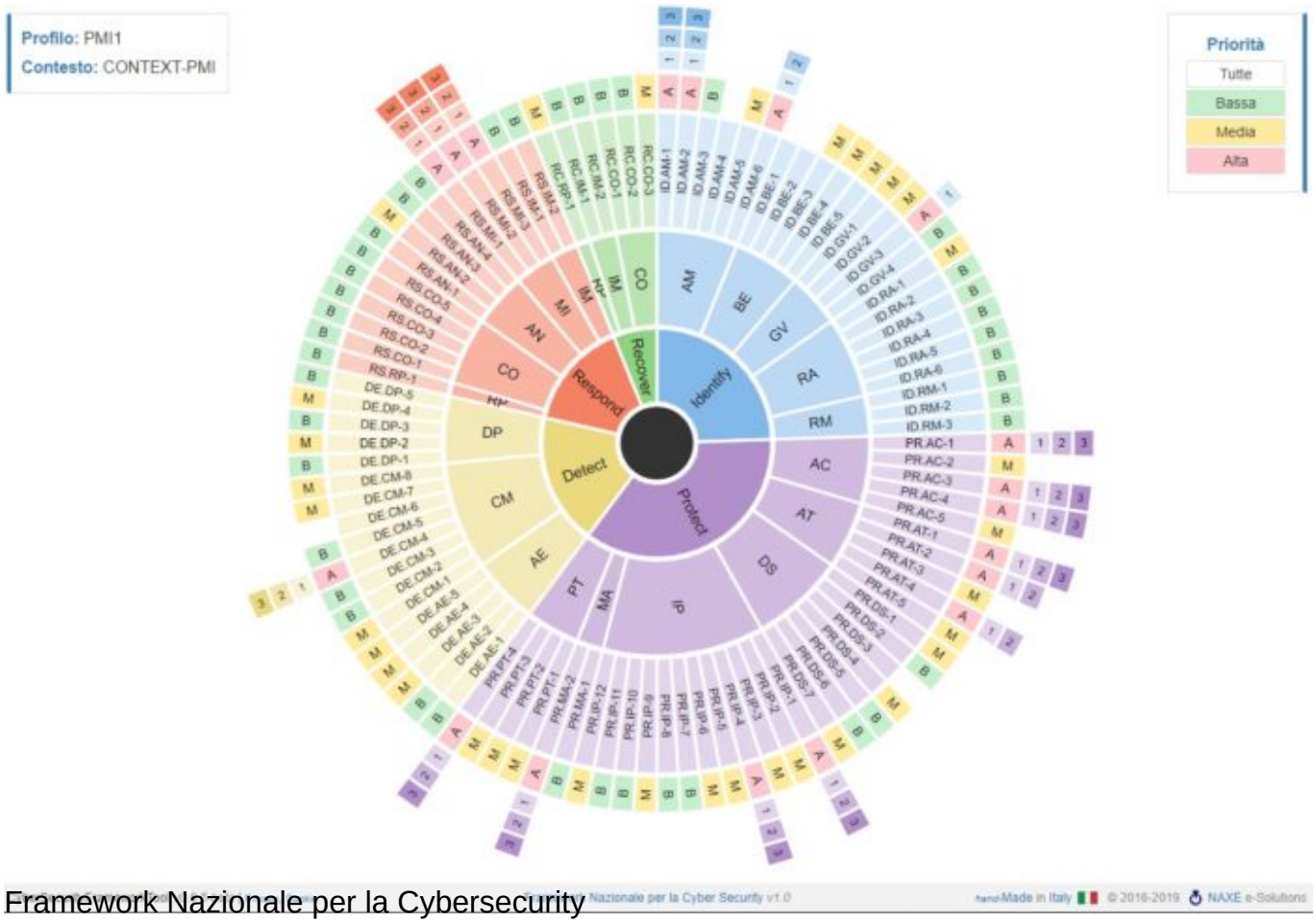
Quesiti che rappresentano sono alcuni dei potenziali "anelli deboli" da considerare nel contesto di una buona analisi dei rischi. Quando si parla di una **buona analisi dei rischi**, occorre valutare, *in primis*, il rischio più grave in relazione a quel particolare contesto, anziché un rischio di minore entità e più facile da mitigare. Di conseguenza, le risorse destinate alla sicurezza dovrebbero essere distribuite in base al risultato di questa analisi e attribuendo un livello di **alta priorità per tutte le criticità evidenziate**.

Altro elemento fondamentale è quello attribuire delle priorità in relazione agli aspetti più critici per poi passare a quelli che in ordine di gravità presentano un livello di rischio inferiore. Lo scopo finale è quello di giungere a una **soglia di rischio accettabile** e sulla base della quale verranno stabilite eventualmente nuove strategie di mitigazione.

Come già detto, nel mondo digitale la sicurezza non è perseguibile in valore assoluto e anche quando tutti gli elementi di rischio, emersi dalla nostra **buona analisi**, siano rientrati nella una soglia di un rischio accettabile, non bisogna mai abbassare la guardia e considerare questi

aspetti in un contesto di insieme e secondo un ciclo di vita continuo e perfettamente integrato nell’infrastruttura tecnologica.

Il Framework Nazionale per la Cybersecurity e la Data Protection ne è un esempio e nasce per supportare le organizzazioni che necessitano di strategie e processi volti alla protezione dei dati personali e alla sicurezza cyber.



Framework Nazionale per la Cybersecurity

A parte tutte le nostre analogie, le **buone pratiche di sicurezza** impongono sempre un approccio di tipo olistico dove la protezione dei componenti nel loro insieme, della catena tecnologica, risulta di gran lunga maggiore della protezione offerta da ogni singolo elemento o singolo anello della stessa.

Gli elementi più deboli saranno comunque sempre presenti in una infrastruttura tecnologica, per quanto evoluta, ma considerando la sicurezza **come un processo e non come un prodotto**, si potranno subito identificare e rafforzare fino al raggiungimento di un livello accettabile del rischio.

La nozione di accettabilità dipende poi dal contesto in cui si opera e da quelle che saranno le scelte strategiche dell’alta direzione.

Stratificare le difese

L’idea alla base di questo principio, di una difesa efficace e profonda rispetto a eventi di rischio, è quella di stratificare le strategie di difesa in modo tale che, se la prima linea dovesse fallire, subentrino subito dopo le altre linee di difesa messe in campo. Torniamo quindi al nostro esempio di sicurezza della banca:

“*perché la nostra banca è più sicura di un piccolo market?*”
Poiché esistono molte misure di sicurezza ridondanti a tutela di un contesto che lo richiede.

Una banca ha per sua natura un prodotto molto ambito, quale il denaro e questo contesto necessita certamente di un’attenzione particolare, che non si limita a un solo elemento di difesa ma si esplica in più misure (strati) a salvaguardia del patrimonio aziendale. Più di questi strati saranno presenti e più saranno le garanzie a tutela della sicurezza. Pertanto, al fianco delle telecamere di sicurezza ci sarà magari un congegno di rilevazione delle presenze, delle procedure di controllo agli ingressi, una guardia che vigila all’esterno e così via.

Come si può facilmente capire, in questo contesto, la superfice di attacco risulta alquanto limitata da tutta una serie di misure di prevenzione che portano a dissuadere gli attaccanti o comunque, far loro capire quali potrebbero essere i rischi e le scarse probabilità di successo anche riuscendo a bypassare le prime linee di difesa. Per esempio, superare l’ingresso o addirittura arrivare agli uffici, non servirebbe a compromettere quello che è il vero patrimonio della banca che invece si troverà dietro l’ultima linea di difesa, all’interno di un più sicuro *caveau*.

Detto questo, comunque, anche avere tutte le misure di sicurezza del caso non garantisce alla banca l’immunità da qualsiasi attacco e che non venga mai derubata o compromessa. Certo è che lo sforzo da mettere in campo, da parte di un attaccante, risulterebbe alquanto difficile e scoraggiante rispetto ad altre realtà (come il piccolo market) dove lo sforzo richiesto sarebbe decisamente a minor costo e con più alte probabilità di successo.

Il principio della **difesa in profondità** o difesa stratificata può sembrare in qualche modo contraddittorio con il principio "***metti in sicurezza l’anello più debole***" perché stiamo essenzialmente dicendo che le difese degli elementi nel loro insieme possono essere più forti dell’elemento più debole.

Come mai invece, non c'è contraddizione?

Il principio del proteggere l’anello più debole si applica quando i componenti hanno un livello di **sicurezza e funzionalità che non si sovrappongono**. Ma quando si tratta di misure di sicurezza ridondanti, la protezione offerta dalla somma delle stesse è di gran lunga maggiore di quella offerta da ogni singolo componente.

I sistemi di ML sono per loro natura costruiti con numerosi componenti e - come più volte evidenziato - i dati, dei quali si nutrono, rappresentano l’elemento più importante da tutelare dal punto di vista della sicurezza.

Ciascuna di queste componenti presenta un certo **fattore di rischio** e, come tale, potenzialmente sfruttabile da attori malintenzionati. In un tale scenario, la vulnerabilità di un singolo componente dovrebbe essere colta da un altro componente, secondo una logica di gioco di squadra. In realtà, nei sistemi complessi, come quelli di AI e ML, non è cosi semplice fare in modo che “tutto” venga gestito e controllato secondo queste logiche di squadra e un mancato controllo a monte potrebbe determinare un problema a valle.

Pensiamo a come la difesa stratificata influisce sull'obiettivo di proteggere i dati di autoapprendimento in un sistema di ML e AI.
Un primo livello di sicurezza tenterà di proteggere i dati di formazione sensibili attraverso qualche tipo di **autenticazione e autorizzazione**, per poi consentire all’utente o al componente così identificato di operare solo dopo che il processo di riscontro dell’identità sia andato a buon fine. Questa potrebbe essere una buona pratica a garanzia della sicurezza ma non di certo sufficiente a garantire che nessuna informazione “sensibile” possa essere divulgata a causa di uso o abuso, dannoso per l'intero sistema. Un elemento di criticità, che non riguarda solamente i sistemi di ML e AI, è certamente dato da un'inadeguata assegnazione dei privilegi a utenti od oggetti che accedono alle informazioni.

Pertanto, un altro aspetto importante da considerare nelle strategie di difesa stratificata e soprattutto su sistemi di ML e AI, riguarda l'applicazione del **principio del privilegio minimo**. Tale pratica consente agli utenti e, in particolare, a coloro che operano con privilegi amministrativi, di poter operare solo ed esclusivamente su determinati ambienti, con determinati strumenti e in relazione a specifici privilegi.

Fuori da questi confini si dovrebbe determinare un’anomalia da notificare ai responsabili della sicurezza per valutarne in tempi rapidi dinamiche e tipologia dell’evento. In questo modo risulta molto più agevole identificare e prevenire gli *exploit* di sicurezza quando ogni componente **limita l'accesso solo alle risorse** effettivamente necessarie.

Identificare e compartimentare i vari componenti di un progetto IT può essere certamente d'aiuto, poiché su ciascuno di questi elementi diventa possibile implementare specifici controlli e politiche di sicurezza secondo quella logica che vede ciascun componente lavorare di concerto con gli altri in un contesto di insieme.

Conclusioni

L’approccio alla sicurezza IT richiede certamente grande consapevolezza da parte di tutti, a partire da coloro che progettano le tecnologie per arrivare a tutti coloro che in qualche modo risultano coinvolti nella filiera tecnologica. È auspicabile che le aziende, con atteggiamento proattivo, mettano in campo strategie e investimenti in grado di mitigare il rischio sulla sicurezza delle informazioni digitali, ma altrettanta attenzione e proattività è richiesta ai governi e alle istituzioni di ogni ordine e grado.

Occorre prendere coscienza che la tutela della sicurezza richiede interventi e regolamentazioni a livello di sistemi paese in grado di fornire garanzie adeguate in termini di fiducia del mercato,

privacy dei cittadini e interessi nazionali.

Riferimenti bibliografici

- Berryville Institute of Machine Learning
- An Architectural Risk Analysis of Machine Learning Systems
- Regolamento europeo per la protezione dei dati personali

Articolo a cura di **Alessandro Bonu**