PuntoSicuro

ARTICOLO DI PUNTOSICURO

Anno 26 - numero 5644 di Mercoledì 19 giugno 2024

Intelligenza artificiale: le indicazioni per difendere i dati personali dal web scraping

Il Garante privacy ha pubblicato le indicazioni per difendere i dati personali pubblicati online da soggetti pubblici e privati in qualità di titolari del trattamento dal web scraping: un'analisi del provvedimento.

Il Garante privacy ha pubblicato con il **provvedimento n. 329 del 20 maggio 2024** le indicazioni per difendere i dati personali pubblicati online da soggetti pubblici e privati in qualità di titolari del trattamento dal web scraping, la raccolta indiscriminata di dati personali su internet, effettuata, da terzi, con lo scopo di addestrare i modelli di Intelligenza artificiale generativa (IAG). Il documento tiene conto dei contributi ricevuti dall'Autorità nell'ambito dell'indagine conoscitiva, deliberata lo scorso dicembre.

Nel documento l'Autorità suggerisce alcune tra le misure concrete da adottare: la creazione di aree riservate, accessibili solo previa registrazione, in modo da sottrarre i dati dalla pubblica disponibilità; l'inserimento di clausole anti-scraping nei termini di servizio dei siti; il monitoraggio del traffico verso le pagine web per individuare eventuali flussi anomali di dati in entrata e in uscita; interventi specifici sui bot utilizzando, tra le altre, le soluzioni tecnologiche rese disponibili dalle stesse società responsabili del web scraping (es: l'intervento sul file robots.txt.).

Si tratta di misure non obbligatorie che i titolari del trattamento dovranno valutare, sulla base del principio di accountability, se mettere in atto per prevenire o mitigare, in maniera selettiva, gli effetti del web scraping, in considerazione di una serie di elementi: lo stato dell'arte tecnologico; i costi di attuazione, in particolare per le PMI.

Con il termine scraping si intende normalmente l'insieme dei meccanismi automatizzati di estrazione delle informazioni da sistemi che per loro caratteristiche tenderebbero a impedirla (o che non sono stati progettati per offrirla).

La forma più comune è il web scraping, cioè l'estrazione di informazioni da siti web. Un sito web pubblico può impedire l'estrazione d'informazione con lo scopo di veicolare la stessa verso l'esterno in maniera controllata (nei tempi e nei modi). Gli strumenti per lo scraping sono di solito script (piccoli programmi) più o meno 'intelligenti' che navigano in rete consultando automaticamente e molto velocemente pagine web e seguendo i link contenuti: durante la navigazione estraggono dati interessanti e li salvano localmente in maniera strutturata e maggiormente usufruibile.

Per esempio, la maggior parte dei servizi di comparazione dei prezzi usa i web scraper per leggere le informazioni sui prezzi di diversi negozi online. Si pensi a quelle piattaforme che forniscono una lista di hotel e voli aerei, comparando i relativi prezzi proposti nei diversi siti web, al fine di permettere all'utente di scegliere il prezzo migliore. Un altro esempio è Google, che effettua abitualmente lo scraping o il "*crawling*" del web per indicizzare i siti web.

Si tratta di una forma di data mining consistente, quindi, nell'utilizzo di un software per estrapolare in maniera automatizzata dati da determinati siti web e nella loro pubblicazione, eventualmente in forma rielaborata, su un altro sito. Il software, programmato per accedere ai dati pubblicati online in maniera sistematica e automatizzata, simula la navigazione di un utente, filtra i dati e li archivia in un database. Esiste anche una forma di scraping manuale, che consiste in un processo di copia e incolla di singole informazioni, utilizzato quando si desidera trovare e memorizzare informazioni mirate, raramente impiegato per grandi quantità di dati a causa dei lunghi tempi di recupero e catalogazione.

Nei casi più semplici, il web scraping può essere effettuato attraverso l'API, detta anche Application Programming Interface di un sito web. Quando un sito web rende disponibile la sua API, chi lavora nello sviluppo web può usarla per estrarre automaticamente dati e altre informazioni utili in un formato conveniente. Ma ovviamente non è sempre così.

In linea di principio il web scraping non è illegale a patto che i dati 'catturati' siano liberamente accessibili sui siti e siano usati per scopi statistici o di monitoraggio dei contenuti. In effetti la maggior parte dei siti web rende i propri dati pubblicamente disponibili a scraper, crawler e altre forme di raccolta automatica dei dati, ma non tutti i dati web sono destinati al pubblico, quindi non tutti i dati web si possono estrarre legalmente. Ma l'aspetto rilevante è dato proprio dall'uso che viene fatto dei dati "*scaricati"* tramite questa attività. Difatti quando si tratta di dati personali e di proprietà intellettuale, il web scraping può trasformarsi rapidamente in web scraping malevolo che si configura anche in altre ipotesi.

Con riferimento alla materia della protezione dei dati personali, quando vengono scaricati dati di natura personali da determinati siti e gli stessi vengano utilizzati in violazione dei principi contenuti nel Regolamento UE sulla protezione dei dati personali n. 679/2016 l'attività dello scraper è sicuramente illecita ed in questi casi bisogna porre particolare attenzione. Si pensi, ad esempio, ad un host web che renda *"accidentalmente*" disponibili al pubblico le informazioni sui propri utenti. Queste potrebbero includere un elenco completo di nomi, email e altre informazioni che tecnicamente sono pubbliche, ma che probabilmente non erano destinate a essere condivise.

Pur essendo tecnicamente legale raccogliere questi dati, non è l'idea migliore. Il fatto che i dati siano pubblici non significa necessariamente che l'host web abbia acconsentito al loro scraping, anche se la sua mancanza di sorveglianza li ha resi pubblici. In tal caso comunque verrebbero violati quei principi di liceità, correttezza, minimizzazione propri della normativa europea. Il Garante privacy già con il provvedimento n. 4 del 14 gennaio 2016 ha inibito ad una società l'utilizzo dei dati personali ? come nomi, cognomi, indirizzi e-mail e numeri di telefonia fissa e cellulare ? di dodici milioni di utenti, che erano stati individuati e raccolti utilizzando lo scraping da diverse pagine web. L'azienda in questione aveva successivamente creato un proprio sito nel quale aveva pubblicato le informazioni raccolte in forma di elenco telefonico online, consultabile anche da altre società per finalità di telemarketing.

Nello stesso modo si è espresso il Garante in un'altra occasione, quando con il provvedimento n. 52 del 01/02/2018 ha vietato a una società di inviare e-mail commerciali a liberi professionisti, i cui indirizzi di posta elettronica e PEC erano stati prelevati da elenchi di pubblico dominio, ma senza chiedere e ottenere la necessaria autorizzazione da parte dei legittimi proprietari. Ma potremmo citare ulteriori casi simili.

In particolare nel provvedimento in esame il Garante suggerisce diverse misure per prevenire o mitigare il web scraping non autorizzato:

1. Creazione di Aree Riservate - La creazione di aree riservate, accessibili solo previa registrazione, è una misura organizzativa volta a limitare la disponibilità pubblica dei dati. Questa pratica consente di sottrarre i dati alla disponibilità indiscriminata, riducendo così le opportunità di web scraping. Tuttavia, tale misura deve essere implementata nel rispetto del principio di minimizzazione dei dati, evitando di richiedere informazioni superflue agli utenti durante la registrazione.

2. Inserimento di Clausole nei Termini di Servizio - L'inserimento di clausole specifiche nei Termini di Servizio (ToS) dei siti web che vietano esplicitamente l'uso di tecniche di web scraping costituisce una misura preventiva di natura giuridica. Queste clausole possono fungere da deterrente legale, consentendo ai gestori dei siti di agire legalmente contro chi non rispetta tali disposizioni.

Ad esempio, piattaforme come YouTube includono nei loro ToS il divieto di accesso tramite mezzi automatizzati senza autorizzazione.

3. Monitoraggio del Traffico di Rete - Il monitoraggio delle richieste HTTP ricevute può aiutare a identificare flussi di dati anomali, indicando possibili attività di scraping. Tecniche come il "Rate Limiting" possono limitare il numero di richieste da indirizzi IP specifici, contribuendo a prevenire attacchi DDoS o scraping eccessivo. Questa è una misura tecnica che può rafforzare la sicurezza dei dati.

4. Intervento sui Bot - Le tecniche di scraping si basano prevalentemente sull'uso di bot. Limitare l'accesso ai bot rappresenta un metodo efficace per contrastare il web scraping. Alcune delle tecniche suggerite includono:

- Verifiche CAPTCHA: Queste verifiche richiedono un'azione umana per procedere, impedendo così l'operatività dei bot.
- Modifica periodica del markup HTML: Cambiare il codice HTML delle pagine web rende più difficile per i bot riconoscere e estrarre i dati.
- Incorporazione dei contenuti in oggetti multimediali: Inserire dati in immagini o altri media rende complessa l'estrazione automatizzata, richiedendo tecnologie di riconoscimento ottico dei caratteri (OCR).

Il Garante sottolinea che, nonostante nessuna delle misure proposte possa impedire completamente il web scraping, esse rappresentano comunque strumenti utili per ridurre i rischi associati alla raccolta non autorizzata di dati personali. È essenziale che i titolari del trattamento valutino attentamente e adottino le misure più adeguate al loro contesto specifico, in conformità con i principi di accountability e protezione dei dati personali previsti dal GDPR.

Il provvedimento del Garante rappresenta senz'altro un passo importante per la protezione dei dati personali nel contesto del web scraping e dell'intelligenza artificiale ma ovviamente da solo non è sufficiente.

L'intervento del Garante ha il grosso merito di promuovere una maggiore consapevolezza tra le aziende che operano nel campo dell'intelligenza artificiale e l'inclusione di clausole specifiche nei Termini di Servizio nonché la possibilità di azioni legali contro i trasgressori rappresentano un forte deterrente per le pratiche di scraping non autorizzato. Anche le raccomandazioni tecniche possono ridurre significativamente l'efficacia del web scraping automatico.

Va però evidenziato che limitare l'accesso ai dati potrebbe rallentare il progresso tecnologico e l'innovazione, specialmente in settori dove l'accesso ai dati pubblici è fondamentale. Inoltre le misure suggerite richiedono risorse tecniche e finanziarie significative, che potrebbero non essere alla portata di tutte le aziende, specialmente le PMI.

D'altro canto la reale efficacia delle misure legali e tecniche dipende dalla capacità di monitorare e far rispettare le normative. I malintenzionati potrebbero trovare modi per aggirare le protezioni, rendendo necessario un aggiornamento continuo delle misure di sicurezza.

Di conseguenza si osserva che:

1. È fondamentale che le misure di prevenzione e mitigazione siano proporzionate e non eccessivamente onerose. Un approccio bilanciato che consenta l'uso responsabile dei dati, senza compromettere la privacy, può essere più sostenibile a lungo termine.
2. Dato che il web scraping e l'uso dei dati avvengono a livello globale, sarebbe utile promuovere una maggiore collaborazione internazionale per stabilire standard e pratiche comuni. Questo aiuterebbe a uniformare le normative e rendere più efficaci le misure di contrasto.
3. Investire in tecnologie avanzate di protezione dei dati, come la crittografia omomorfica o le tecniche di privacy differenziale, potrebbe offrire nuove opportunità per proteggere i dati personali senza limitarne l'uso per l'addestramento di modelli di intelligenza artificiale.

Michele Iaselli

Fonte: federprivacy

---

www.puntosicuro.it