maggio 2024

Web scraping ed intelligenza artificiale generativa:

nota informativa e possibili azioni di contrasto

Introduzione

Con il presente documento il Garante intende fornire prime indicazioni sul fenomeno della raccolta massiva di dati personali dal web per finalità di addestramento dei modelli di intelligenza artificiale generativa (di seguito anche "IAG") e segnalare possibili azioni di contrasto che i gestori di siti internet e di piattaforme online, sia pubblici che privati, operanti in Italia, quali titolari del trattamento dei dati personali oggetto di pubblicazione, potrebbero implementare al fine di prevenire, ove ritenuta incompatibile con le basi giuridiche e le finalità della pubblicazione, la raccolta di dati da parte di terzi per finalità di addestramento dei modelli di intelligenza artificiale.

Il presente documento concerne esclusivamente dati personali oggetto di diffusione in quanto pubblicati su siti web e piattaforme online.

Il documento tiene conto dei contributi ricevuti dall'Autorità nell'ambito dell'indagine conoscitiva in materia di *web scraping*, deliberata con provvedimento del 21 dicembre 2023, pubblicato nella Gazzetta Ufficiale n. 14 del 18 gennaio 2024.

Ad ogni modo sono rimesse ai gestori dei suddetti siti e piattaforme, pubblici e privati, nella misura in cui siano al contempo titolari del trattamento dei dati personali ai sensi del Regolamento (UE) 2016/679 (di seguito "RGPD"), le valutazioni da effettuare caso per caso, sulla base della natura, dell'ambito di applicazione, del contesto e delle finalità dei dati personali trattati, del regime di pubblicità, accesso e riuso da assicurare, della tutela apprestata da altre specifiche normative (ad esempio, la normativa a tutela del diritto di autore), tenendo conto dello stato dell'arte (inteso in senso precipuamente tecnologico) e dei costi di attuazione (in particolare con riferimento alle piccole e medie imprese).

Web scraping e diritto alla protezione dei dati personali

Nella misura in cui il *web scraping* implica la raccolta di informazioni riconducibile a una persona fisica indentificata o identificabile si pone un problema di protezione dati personali.

Il focus della *compliance* con il RGPD viene generalmente puntato sui soggetti che trattano i dati personali raccolti tramite tecniche di *web scraping*, in particolare con riferimento all'individuazione di una idonea base giuridica ai sensi dell'art. 6 del RGPD per la trattazione di tali dati[1], la cui individuazione deve essere effettuata sulla base di una valutazione di idoneità che il titolare deve essere in grado di comprovare, in base al principio di *accountability* di cui all'art. 5, par. 2, RGPD.

Questo documento propone una diversa prospettiva, esaminando la posizione dei soggetti, pubblici e privati, gestori di siti *web* e piattaforme *online,* operanti quali titolari del trattamento di dati personali, che rendano pubblicamente disponibili, dati (anche personali) che vengono raccolti dai *bot* di terze parti.

[1] Il Garante ha, in passato, dichiarato illecita l'attività di *web scraping* posta in essere dalla società statunitense Clearview, [doc web n. 9751362], reperibile all'URL https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9751362 e quella effettuata dalla piattaforma Trovanumeri [doc web n. 9903067], reperibile all'URL https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9903067.

In linea con tale impostazione, il documento indica alcune tra le possibili cautele che, sulla scorta di una valutazione da effettuarsi caso per caso, i titolari del trattamento di dati personali resi disponibili *online* per finalità diverse e sulla base di differenti condizioni di legittimità possono implementare al fine di prevenire o mitigare, in maniera selettiva, l'attività di *web scraping* per finalità di addestramento di modelli di intelligenza artificiale generativa.
Al riguardo pare opportuno ricordare che ogni titolare del trattamento di dati personali, soggetto pubblico o privato, ai sensi del Regolamento può rendere disponibili al pubblico tali dati personali esclusivamente per finalità specifiche e sulla base di una o più condizioni di legittimità tra quelle previste all'art. 6 del Regolamento (es: obblighi di trasparenza, pubblicità legale, procedure a evidenza pubblica, diritto di cronaca, contratto in essere con gli interessati).

Il giudizio di liceità del *web scraping* deve, dunque, essere effettuato caso per caso sulla base dei diversi e contrapposti diritti in gioco: in tal senso, per le finalità di questo documento, tale liceità non è e non può che essere oggetto di valutazione in termini meramente teorici.
Si precisa, inoltre, che il presente documento non si occupa di indicare le misure di sicurezza che i titolari del trattamento debbono implementare per proteggere i dati personali da operazioni qualificabili come *web scraping* "malevolo", in quanto in grado di sfruttare delle vulnerabilità dei sistemi informativi non adeguatamente protetti dal punto di vista della sicurezza informatica. Sotto tale profilo rimane fermo, ai sensi dell'art. 32 del RGPD, l'obbligo in capo ai titolari del trattamento di assicurare, su base permanente, la riservatezza, l'integrità, la disponibilità e la resilienza dei sistemi e dei servizi di trattamento. A tal proposito, si richiamano i principi espressi nella decisione adottata, nel novembre 2022, dall'autorità irlandese nei confronti di Meta Platforms Ireland Ltd[2] in merito alla mancata adeguata protezione dei dati (a causa di impostazioni non conformi al RGPD degli strumenti Facebook Search, Facebook Messenger *Contact Importer* e *Instagram Contact Importer*) ed alla conseguente raccolta *online*, tramite tecniche di *web scraping* adottate da terze parti, dei dati di circa 533 milioni di utenti del servizio Facebook nel periodo compreso tra il 25 maggio 2018 e settembre 2019.[3]

Le tecniche di raccolta massiva di dati dal web e le loro finalità

La nascita e l'affermazione di Internet sono intrinsecamente connesse alla sua architettura tecnologica aperta basata su standard informatici *de facto,* indipendenti da specifiche "proprietarie", fondati sulla *suite* di protocolli TCP (*Transmission Control Protocol*) e IP (*Internet Protocol*). Con il tempo, a tali protocolli si è aggiunto, tra gli altri, il protocollo HTTP (*Hyper Text Transfer Protocol*) con il quale, a seguito della decisione del CERN di Ginevra di renderlo pubblico nel 1990, è stato possibile lo sviluppo libero del *World Wide Web* (di seguito "web") così come lo

[2] https://www.dataprotection.ie/sites/default/files/uploads/2022-12/Final%20Decision_IN-21-4-2_Redacted.pdf.
[3] Il data breach era stato attenzionato al pubblico anche dal Garante mediante l'adozione di un provvedimento generale di avvertimento rivolto a tutte le persone fisiche o giuridiche, le autorità pubbliche, i servizi e qualsiasi organismo che, singolarmente o insieme ad altri svolgeva nell'ambito dei trattamenti di dati personali il ruolo di titolari o di responsabili del trattamento. Il provvedimento chiariva che eventuali trattamenti dei dati personali oggetto del *data breach* occorso a Meta, si sarebbero posti in violazione degli artt. 5, par. 1, lett. a), 6 e 9 del Regolamento, con tutte le conseguenze, anche di carattere sanzionatorio, previste dalla disciplina in materia di protezione dei dati personali [doc web 9574600]. Reperibile all'URL https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9574600.

conosciamo, con la prima formalizzazione in forma di standard (HTTP/1.1) con il documento RFC-2068 del 1997.

La navigazione nel *web* si basa, quindi, su protocolli aperti che consentono di reperire informazioni e dati pubblicamente disponibili *online* oppure resi disponibili in aree ad accesso controllato. Informazioni e dati possono essere raccolti in maniera sistematica anche attraverso programmi (*web robot* o, più semplicemente, *bot*) che operano in maniera automatizzata simulando la navigazione umana, a condizione che le risorse (*e.g.* siti *web*, contenuti, etc.) visitate da questi ultimi risultino accessibili al pubblico indistinto e non sottoposte a controlli di accesso.

Un recente studio condotto da Imperva,[4] una società del gruppo francese Thales, ha rivelato che, nell'anno 2023, il 49,6% di tutto il traffico Internet è stato generato dai *bot* con un aumento pari al 2,1% rispetto all'anno precedente, aumento che è stato parzialmente ricondotto alla diffusione di sistemi di intelligenza artificiale e, in particolare, dei modelli linguistici di grandi dimensioni (di seguito anche "LLM" - *Large Language Model*) sottesi all'intelligenza artificiale generativa.[5]

Nell'ambiente *online* i più noti *bot* utilizzati sono i *web crawler* (detti anche "*spider*") dei motori di ricerca. Si tratta di programmi che scandiscono sistematicamente il *web* al fine di raccogliere i dati contenuti nelle pagine *web* ed indicizzarli per garantire il funzionamento dei motori di ricerca (GoogleBot e BingBot, ad esempio, sono gli *spider* dei motori di ricerca di Google e di Microsoft).

Si parla di *web scraping* laddove l'attività di raccolta massiva ed indiscriminata di dati (anche personali) condotta attraverso tecniche di *web crawling* è combinata con un'attività consistente nella memorizzazione e conservazione dei dati raccolti dai *bot* per successive mirate analisi, elaborazioni ed utilizzi.[6]

Le finalità per cui vengono impiegati i *bot* e svolta attività di *web scraping sono molteplici*, talune sono senz'altro malevoli (si pensi ai tradizionali attacchi DDoS - *Distributed Denial of Service* – ai tentativi di *login* forzato, allo *scalping*, al furto di credenziali ed alle frodi digitali), mentre per tali altre la valutazione di liceità o illiceità resta inevitabilmente rimessa a un accertamento da compiersi caso per caso sulla base di una pluralità di valutazioni di competenza sotto taluni profili del soggetto che vi procede e sotto taluni altri al soggetto che pubblica i dati personali che formano oggetto di tale attività. Tra le finalità alla base dell'attività di web scraping, come si è anticipato, vi è anche quella di addestramento di algoritmi di intelligenza artificiale generativa[7]. I grandi *dataset* utilizzati dagli sviluppatori di intelligenza artificiale generativa hanno provenienze variegate, ma il *web scraping* costituisce un denominatore comune. Gli sviluppatori possono, infatti, utilizzare *dataset* oggetto di autonoma attività di *scraping*, oppure attingere da *data lake* di terze parti (tra questi si menzionano, a titolo soltanto esemplificativo, l'*open repository* della *non*-profit statunitense Common Crawl,[8] i dataset della piattaforma franco-americana Hugging Face[9] o della non-profit

[4] https://www.imperva.com/resources/resource-library/reports/2024-bad-bot-report/

[5] Per dare un'idea del fenomeno, si rappresenta che dieci anni or sono, nel 2013, il *traffic* Internet *traffic* era costituito al 23.6% da traffico generato da bot cattivi (*bad bot*), al 19.4% da bot buoni (good bot) e al 57% da umani.

[6] Ai fini di questo documento si utilizzerà il termine *web scraping* come comprensivo anche del *web crawling*.

[7] Si intende intelligenza artificiale generativa un sistema di intelligenza artificiale in grado di generare nuovi testi, immagino, audio e video.

[8] https://commoncrawl.org/.

[9] https://huggingface.co/.

tedesca LAION AI[10]) i quali sono stati, a loro volta, precedentemente creati mediante operazioni di *scraping*. Per contro, è possibile che i dataset di addestramento siano costituiti dai dati già in possesso degli sviluppatori, come ad esempio i dati degli utenti di servizi offerti dal medesimo sviluppatore o i dati degli utenti di un social network.

Possibili azioni di contrasto al *web scraping* per finalità di addestramento dell'intelligenza artificiale generativa

Al netto, dunque, degli obblighi attualmente gravanti sui titolari del trattamento connessi sia ai regimi di pubblicità, accesso e riuso dei dati previsti *ex lege* che alle misure di sicurezza necessarie per garantire la protezione dei dati, il Garante ritiene utile fornire alcune indicazioni ai gestori dei siti web e di piattaforme online, operanti in Italia quali titolari del trattamento di dati personali resi disponibili al pubblico attraverso piattaforme online, in merito alle possibili cautele che potrebbero essere adottate per mitigare gli effetti del *web scraping* di terze parti, finalizzato all'addestramento di sistemi di intelligenza artificiale generativa ove considerato, in attuazione del principio di accountability dal singolo titolare del trattamento, incompatibile con le finalità e le basi giuridiche della messa a disposizione del pubblico dei dati personali.

Nella piena consapevolezza che nessuna di tali misure può ritenersi idonea a impedire al 100% il *web scraping, esse devono considerarsi* cautele da adottarsi sulla base di un'autonoma valutazione del titolare del trattamento, in attuazione del principio di responsabilizzazione (*accountability*), allo scopo di impedire l'utilizzazione ritenuta non autorizzata, da parte di terzi, dei dati personali pubblicati in qualità di titolare.

1. <u>Creazione di aree riservate</u>

Atteso che l'addestramento dell'intelligenza artificiale generativa si basa su enormi quantità di dati che spesso provengono da attività di *web scraping* diretta (ovverosia effettuata dallo stesso soggetto che sviluppa il modello), indiretta (ovverosia effettuata su dataset creati mediante tecniche di *web scraping* da soggetti terzi rispetto allo sviluppatore del modello) od ibrida, su fonti presenti nel web, la creazione di aree riservate, a cui si può accedere solo previa registrazione, rappresenta una valida cautela in quanto sottrae dati dalla ritenuta pubblica disponibilità. Tale tipologia di cautela tecnico-organizzativa può, sebbene indirettamente contribuire ad una maggiore tutela dei dati personali rispetto ad attività di web scraping.

Di contro, tale misura non può dar luogo ad un trattamento di dati eccessivo da parte del titolare, in violazione del principio di minimizzazione di cui all'articolo 5, par. 1, lett. c), RGPD (a titolo esemplificativo, si ricorda che i titolari del trattamento non dovrebbero imporre in sede di registrazione, agli utenti che navighino sui loro siti *web* o sulle loro piattaforme *online* e che fruiscano dei relativi servizi, oneri di registrazione ulteriori ed ingiustificati [11].

[10] https://laion.ai/.

[11] Si richiama, in tal senso, una recente decisione, adottata nell'ambito della procedura di cooperazione europea *ex* art.60 ss RGPD, con cui l'autorità finlandese ha sostenuto l'illiceità dell'obbligo imposto dal titolare del trattamento di creare un account utente per il perfezionamento di un singolo acquisto online su una piattaforma di e-commerce. Reperibile all'URL https://tietosuoja.fi/en/-/administrative-fine-imposed-on-verkkokauppa.com-for-failing-to-define-storage-period-of-customer-data-requiring-customers-to-register-was-also-illegal.

2. Inserimento di clausole *ad hoc* nei termini di servizio

L'inserimento nei Termini di Servizio (ToS) di un sito *web* o di una piattaforma *online* dell'espresso divieto di utilizzare tecniche di *web scraping* costituisce una clausola contrattuale che, se non rispettata, consente ai gestori di detti siti e piattaforme di agire in giudizio per far dichiarare l'inadempimento contrattuale della controparte. Si tratta di una cautela di mera natura giuridica che opera, in quanto tale *ex post*, ma che può fungere da strumento di carattere special-preventivo e, in tal modo, fungere da deterrente, contribuendo ad una maggiore tutela dei dati personali rispetto ad attività di *web scraping*. A tal proposito, si richiamano l'ampio utilizzo e l'efficacia di tale misura, in particolare, nella protezione dei contenuti protetti dal diritto d'autore (si menzionano, tra i tanti, i termini di servizio di YouTube, a cui Google vieta l'accesso con mezzi automatizzati, quali robot, botnet o strumenti di *scraping*, salvo si tratti di motori di ricerca pubblici, in conformità con il file robots.txt di YouTube o salvo previa autorizzazione scritta da parte di YouTube[12]).

3. Monitoraggio del traffico di rete

Un semplice accorgimento tecnico quale il monitoraggio delle richieste HTTP ricevute da un sito web o da una piattaforma consente di individuare eventuali flussi anomali di dati in ingresso ed in uscita da un sito web o da una piattaforma online e di intraprendere adeguate contromisure di protezione. Tale cautela può essere accompagnata anche da un *Rate Limiting*, una misura tecnica che permette di limitare il traffico di rete ed il numero di richieste selezionando solo quelle provenienti da determinati indirizzi IP, al fine di impedire *a priori* un traffico eccessivo di dati (in particolare attacchi DDoS o *web scraping*). Si tratta di cautele di natura tecnica che, sebbene indirettamente, possono contribuire ad una maggiore tutela dei dati personali rispetto ad attività di *web scraping* per finalità di addestramento dell'intelligenza artificiale generativa.

4. Intervento sui bot

Come sopra illustrato, il *web scraping* si basa sull'utilizzo di bot. Qualunque tecnica in grado di limitare l'accesso ai *bot* si rivela, pertanto, un efficace metodo per arginare l'attività automatizzata di raccolta dati che viene effettuata tramite tali software. È doveroso sottolineare che nessuna tecnica che agisce sui *bot* è in grado di annullarne l'operatività al 100%, ma anche che alcune azioni di contrasto possono indubbiamente contribuire a prevenire/mitigare il *web scraping* non desiderato per finalità di addestramento dell'intelligenza artificiale generativa.

A tal proposito si menzionano, a titolo meramente esemplificativo:

i) l'inserimento di verifiche CAPTCHA (*Completely Automated Public Turing-test-to-tell Computers and Humans Apart*) le quali, imponendo un'azione eseguibile solo da un essere umano, impediscono l'operatività dei bot;
ii) la modifica periodica del *markup* HTML, in modo da ostacolare o comunque rendere più complicato lo *scraping* da parte dei *bot*. Tale modifica può essere realizzata mediante annidamento di elementi HTML oppure modificando altri aspetti del *markup*, anche in maniera randomica.
iii) l'incorporazione dei contenuti ovvero dei dati che si intendono sottrarre alle attività di *scraping* all'interno di oggetti multimediali, quali ad esempio immagini (si pensi all'uso di

[12] https://www.youtube.com/t/terms#6bedad2de4.

tale tecnica nel caso di testo breve come numeri di telefono o *email*) o altre forme di media. In questo caso l'estrazione dei dati da parte del *bot* risulterebbe significativamente più complessa. Ad esempio, per l'estrazione dei dati dall'immagine – posto che il *bot* sia stato in grado di identificarne la presenza ivi codificata – sarebbe necessario il riconoscimento ottico dei caratteri (OCR), non esistendo il contenuto come stringa di caratteri nel codice della pagina *web*. Corre tuttavia segnalare come una tal misura, pur rappresentando una possibile forma di sottrazione di alcuni dati all'attività di *scraping*, potrebbe rappresentare un ostacolo per gli utenti che perseguano alcuni legittimi fini, (*e.g.* impossibilità di copiare i contenuti dal sito *web*).

iv) il monitoraggio dei *file* di *log,* al fine di bloccare eventuali *user-agent* non desiderati, ove identificabili[13];

v) l'intervento sul *file* robot.txt. Il *file* robot.txt è uno strumento tecnico che, dal giugno 1994, riveste un ruolo fondamentale nella gestione dell'accesso ai dati contenuti nei siti web, in quanto consente ai gestori di indicare se l'intero sito o alcune sue parti possono o meno essere oggetto di indicizzazione e *scraping*. Creato come strumento per regolare l'accesso dei *crawler* dei motori di ricerca (e quindi per controllare l'indicizzazione dei siti web) l'accorgimento basato sul *robots.txt* (sostanzialmente, una *black-list* di contenuti da sottrarre all'indicizzazione) si è evoluto nel REP (*Robot Exclusion Protocol*), un protocollo informale per consentire (*allow*) o non consentire (*disallow*) l'accesso alle diverse tipologie di bot. Nel caso di specie, è teoricamente ipotizzabile l'inserimento nel *file* robot.txt di indicazioni volte a non consentire (*disallow*) l'azione di specifici *bot* finalizzati allo *scraping* per finalità di addestramento dell'intelligenza artificiale generativa facenti capo a determinati sviluppatori. Esistono, infatti, alcuni *bot* che, per autodichiarazione degli stessi sviluppatori di IAG, sono finalizzati allo *scraping* per tali finalità. Si riportano, a titolo meramente esemplificativo, i *bot* di OpenAI (GPTBot)[14] e di Google (Google-Extended)[15], che possono essere esclusi, tramite REP, per prevenire lo *scraping* totale o parziale di un sito *web* da parte dei relativi sviluppatori. Si tratta di una misura tecnica mirata, ma limitata nella sua efficacia per diversi ordini di motivi, tra cui:1) il REP non è uno *standard* riconosciuto e, pertanto, il suo rispetto si basa solo sull'assunzione di un impegno etico da parte dei *web scraper*; 2) esistono *bot* che raccolgono dati dal *web* mediante tecniche di *scraping* per finalità non esclusivamente di addestramento di IAG ed ai cui *data lake* gli sviluppatori di IAG ricorrono frequentemente per le proprie finalità (tra questi, il più noto è sicuramente il CCBot della non-profit Common Crawl, sopra citata); 3) similmente, esistono *bot* di sviluppatori di IAG la cui finalità non è stata esplicitamente dichiarata o di cui non sono stati condivisi i dettagli tecnici, per cui è difficile conoscere i comportamenti e gli scopi del loro utilizzo (*e.g.* ClaudeBot di Anthropic).

[13] Gli *user-agent* possono anche essere anonimi o indicare un nome non qualificante o essere oggetto di *spoofing*.
[14] https://platform.openai.com/docs/gptbot.
[15] https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers?hl=it. *Google-Extended* è diverso dal *crawler* principale di Google (*Googlebot*) che viene utilizzato per il funzionamento del motore di ricerca di Google e non influisce sull'inserimento o sul *ranking* di un sito in detto motore.

Conclusione

L'intelligenza artificiale generativa è foriera di benefici per la collettività che non possono essere limitati, negati, né sminuiti. L'addestramento dei modelli sottesi al funzionamento di tali sistemi richiede, tuttavia, una mole ingente di dati (anche di carattere personale), spesso provenienti da una raccolta massiva ed indiscriminata effettuata sul *web* con tecniche di *web scraping*. I gestori di siti *web* e di piattaforme *online* che rivestano al tempo stesso il ruolo di titolari del trattamento, fermi restando gli obblighi di pubblicità, accesso, riuso e di adozione delle misure di sicurezza previste dal RGPD, dovrebbero valutare, caso per caso, quando risulti necessario, in conformità alla vigente disciplina, sottrarre i dati personali che trattano ai *bot* di terze parti mediante l'adozione di azioni di contrasto come quelle indicate che, sebbene non esaustive né per metodo, né per risultato, possono contenere gli effetti dello *scraping finalizzato all'addestramento degli algoritmi di intelligenza artificiale generativa*.