GARANTE PER LA PROTEZIONE DEI DATI PERSONALI

DELIBERA 20 maggio 2024

Nota informativa in materia di web scraping per finalita' di addestramento di intelligenza artificiale generativa e di possibili azioni di contrasto a tutela dei dati personali. (Provvedimento n. 329). (24A02916)

(GU n.132 del 7-6-2024)

IL GARANTE PER LA PROTEZIONE
DEI DATI PERSONALI

Nella riunione odierna, alla quale hanno preso parte il prof. Pasquale Stanzione, presidente, la prof.ssa Ginevra Cerrina Feroni, vicepresidente, il dott. Agostino Ghiglia e l'avv. Guido Scorza, componenti, e il cons. Fabio Mattei, segretario generale;

Visto il regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonche' alla libera circolazione di tali dati (regolamento generale sulla protezione dei dati, di seguito, «regolamento»);

Visto il codice in materia di protezione dei dati personali, recante disposizioni per l'adeguamento dell'ordinamento nazionale al regolamento (UE) 2016/679 (decreto legislativo 30 giugno 2003, n. 196, come modificato dal decreto legislativo 10 agosto 2018, n. 101, di seguito «codice»);

Visto l'art. 57, par. 1, lettera b), del regolamento che attribuisce alle autorita' di controllo il compito di promuovere la consapevolezza e favorire la comprensione del pubblico riguardo ai rischi, alle norme, alle garanzie e ai diritti in materia di protezione dei dati personali, con particolare attenzione alle attivita' destinate specificamente ai minori;

Rilevato che e' nota l'attivita' di raccolta massiva, attraverso tecniche di web scraping, da parte di soggetti che sviluppano sistemi di intelligenza artificiale generativa (di seguito anche «IAG») di grandi quantita' di dati, anche personali, pubblicati nei siti web e nelle piattaforme online gestiti da soggetti pubblici e privati, stabiliti nel territorio italiano, per finalita' di addestramento di tale tipologia intelligenza artificiale;

Considerato che restano fermi in capo ai soggetti che trattano dati personali, direttamente od indirettamente raccolti tramite tecniche di web scraping per finalita' di addestramento di IAG, nella misura in cui essi siano qualificabili come titolari del trattamento, tutti gli obblighi previsti dal regolamento, in particolare l'obbligo di individuare una idonea base giuridica per il trattamento di tali dati, in ossequio al principio di accountability di cui all'art. 5, par. 2, del regolamento;

Considerato che restano ferme le disposizioni in materia di obblighi di pubblicazione per finalita' di trasparenza di cui al decreto legislativo n. 33/2013 e altre pubblicita' legali, in materia di apertura dei dati e riutilizzo dell'informazione del settore pubblico ai sensi decreto legislativo n. 36/2006 (successive modificazioni ed integrazioni), in materia di prevenzione della corruzione e trasparenza da parte di societa' ed enti di diritto privato controllati e partecipati dalle pubbliche amministrazioni e di enti pubblici economici, nonche' le disposizioni previste da normative specifiche come quelle a tutela della proprieta' intellettuale e del diritto d'autore;

Considerato che restano fermi altresi', in quanto applicabili, i regolamenti (cosiddetti Digital Markets Act, Digitale Service Act, Digital Governance Act e Data Act) che hanno dato attuazione alla strategia digitale europea finalizzata alla creazione di uno spazio comune europeo dei dati in grado di stimolare l'innovazione digitale nell'ambito del mercato unico europeo;

Considerato che grava sui gestori, pubblici e privati, di siti web e piattaforme online, stabiliti in Italia, l'onere di rispettare i principi fondamentali previsti dal regolamento e, in particolare, di quelli di limitazione della finalita', di minimizzazione dei dati e di integrita' e riservatezza, ai sensi dell'art. 5, par. 1, lettera b), c) ed f) del regolamento;

Tenuto conto delle peculiarita' dell'innovazione tecnologica sottesa ai sistemi di IAG nella misura in cui essa comporta dei rischi per la protezione dei dati personali, in particolare con riferimento alle difficolta' in capo agli interessati di esercitare in maniera efficace i diritti di cui agli articoli da 13 a 22 del regolamento;

Tenuto conto dei contributi ricevuti dall'Autorita' 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;

Considerato che taluni soggetti che sviluppano sistemi di intelligenza artificiale generativa hanno essi stessi messo a disposizione dei gestori, pubblici e privati, di siti web e piattaforme online misure tecnologiche che consentono di escludere, in tutto o in parte, il contenuto dei loro siti e piattaforme dall'attivita' web scraping indesiderato;

Considerato che la pubblicazione di dati personali online da parte di qualunque titolare pubblico o privato di dati personali deve essere sempre basata su una o piu' basi giuridiche tra quelle indicate all'art. 6 del regolamento e deve sempre avvenire per una o piu' specifiche finalita' normalmente diverse dalla costituzione di basi di dati aperte utilizzabili da terzi per l'addestramento di algoritmi di intelligenza artificiale;

Ritenuto utile fornire prime indicazioni sulle tecniche di raccolta massiva di dati personali dal web per finalita' di addestramento dei modelli di intelligenza artificiale generativa ed indicare ai gestori dei siti web e di piattaforme online, stabiliti in Italia, nella misura in cui siano qualificabili come titolari del trattamento, possibili azioni di contrasto che potrebbero essere adottate per mitigare gli effetti del web scraping di terze parti finalizzato all'addestramento di sistemi di intelligenza artificiale generativa che venisse ritenuto incompatibile, dal singolo titolare del trattamento, in attuazione del principio di accountability, con le finalita' e la base giuridica della pubblicazione dei dati personali nonche' con eventuali disposizioni di legge speciali vigenti di terze parti finalizzato all'addestramento di sistemi di intelligenza artificiale generativa;

Rilevato che dette misure non debbono essere intese come obblighi di messa in conformita' al regolamento bensi' come misure tecniche ed organizzative la cui adozione dovra' essere valutata tenuto conto anche dello stato dell'arte e dei costi di attuazione (in particolare con riferimento alle PMI), nonche' della natura, dell'ambito di applicazione, del contesto e delle finalita' dei trattamenti effettuati e tenuto altresi' conto che si tratta di misure non esaustive, da un punto di vista tecnologico, rispetto al fenomeno del web scraping indesiderato;

Viste le osservazioni formulate dal segretario generale ai sensi dell'art. 15 del regolamento n. 1/2000;

Relatore il prof. Pasquale Stanzione;

Delibera:

1) Ai sensi dell'art. 57, par. 1, lettera b), del regolamento, di adottare la «Nota informativa» in materia di web scraping per finalita' di addestramento di intelligenza artificiale generativa e di possibili azioni di contrasto a tutela dei dati personali» contenuta nel documento allegato che forma parte integrante della presente deliberazione.

2) ai sensi dell'art. 154-bis, comma 3, del codice, che copia del presente provvedimento sia trasmessa al Ministero della giustizia per la sua pubblicazione nella Gazzetta Ufficiale della Repubblica italiana, a cura dell'Ufficio competente.

Il presidente e relatore
Stanzione

Il segretario generale
Mattei

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 finalita' 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 finalita' della pubblicazione, la raccolta di dati da parte di terzi per finalita' 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'Autorita' 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 finalita' dei dati personali trattati, del regime di pubblicita', 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 idoneita' 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.

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 finalita' diverse e sulla base di differenti condizioni di legittimita' possono implementare al fine di prevenire o mitigare, in maniera selettiva, l'attivita' di web scraping per finalita' 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 puo' rendere disponibili al pubblico tali dati personali esclusivamente per finalita' specifiche e sulla base di una o piu' condizioni di legittimita' tra quelle previste all'art. 6 del regolamento (es: obblighi di trasparenza, pubblicita' legale, procedure a evidenza pubblica, diritto di cronaca, contratto in essere con gli interessati).

Il giudizio di liceita' del web scraping deve, dunque, essere effettuato caso per caso sulla base dei diversi e contrapposti diritti in gioco: in tal senso, per le finalita' di questo documento, tale liceita' non e' e non puo' 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 vulnerabilita' 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'integrita', la disponibilita' 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'autorita' 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 finalita'.

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 e' 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, e' stato possibile lo sviluppo libero del World Wide Web (di seguito «web») cosi' come lo 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, piu' 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 societa' del gruppo francese Thales, ha rivelato che, nell'anno 2023, il 49,6% di tutto il traffico internet e' stato generato dai bot con un aumento pari al 2,1% rispetto all'anno precedente, aumento che e' 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 piu' 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'attivita' di raccolta massiva ed indiscriminata di dati (anche personali) condotta attraverso tecniche di web crawling e' combinata con un'attivita' consistente nella memorizzazione e conservazione dei dati raccolti dai bot per successive mirate analisi, elaborazioni ed utilizzi (6) .

Le finalita' per cui vengono impiegati i bot e svolta attivita' 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 liceita' o illiceita' resta inevitabilmente rimessa a un accertamento da compiersi caso per caso sulla base di una pluralita' 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 attivita'. Tra le finalita' alla base dell'attivita' di web scraping, come si e' anticipato, vi e' 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 attivita' 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 tedesca LAION AI (10)) i quali sono stati, a loro volta, precedentemente creati mediante operazioni di scraping. Per contro, e' possibile che i dataset di addestramento siano costituiti dai dati gia' 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 finalita' di addestramento dell'intelligenza artificiale generativa.

Al netto, dunque, degli obblighi attualmente gravanti sui titolari del trattamento connessi sia ai regimi di pubblicita', 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 finalita' e le basi giuridiche della messa a disposizione del

pubblico dei dati personali.

Nella piena consapevolezza che nessuna di tali misure puo' 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 qualita' di titolare.

1. Creazione di aree riservate.

Atteso che l'addestramento dell'intelligenza artificiale generativa si basa su enormi quantita' di dati che spesso provengono da attivita' 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 puo' accedere solo previa registrazione, rappresenta una valida cautela in quanto sottrae dati dalla ritenuta pubblica disponibilita'. Tale tipologia di cautela tecnico-organizzativa puo', sebbene indirettamente contribuire ad una maggiore tutela dei dati personali rispetto ad attivita' di web scraping.

Di contro, tale misura non puo' dar luogo ad un trattamento di dati eccessivo da parte del titolare, in violazione del principio di minimizzazione di cui all'art. 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) .

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 puo' fungere da strumento di carattere special-preventivo e, in tal modo, fungere da deterrente, contribuendo ad una maggiore tutela dei dati personali rispetto ad attivita' 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 conformita' 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 puo' 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 attivita' di web scraping per finalita' 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'attivita' automatizzata di raccolta dati che viene effettuata tramite tali software. E' doveroso sottolineare che nessuna tecnica che agisce sui bot e' in grado di annullarne l'operativita' al 100%, ma anche che alcune azioni di contrasto possono indubbiamente contribuire a prevenire/mitigare il web scraping non desiderato per finalita' 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'operativita' dei bot;

ii) la modifica periodica del markup HTML, in modo da ostacolare o comunque rendere piu' complicato lo scraping da parte dei bot. Tale modifica puo' 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 attivita' di scraping all'interno di oggetti multimediali, quali ad esempio immagini (si pensi all'uso di 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 piu' 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'attivita' di scraping, potrebbe rappresentare un ostacolo per gli utenti che perseguano alcuni legittimi fini, (e.g. impossibilita' di copiare i contenuti dal sito web).

iv) il monitoraggio dei file di log, al fine di bloccare eventuali user-agent non desiderati, ove identificabilin (13) ;

v) l'intervento sul file robot.txt. Il file robot.txt e' 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 e' 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, e' teoricamente ipotizzabile l'inserimento nel file robot.txt di indicazioni volte a non consentire (disallow) l'azione di specifici bot finalizzati allo scraping per finalita' 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 finalita'. 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 e' 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 finalita' non esclusivamente di addestramento di IAG ed ai cui data lake gli sviluppatori di IAG ricorrono frequentemente per le proprie finalita' (tra questi, il piu' noto e' sicuramente il CCBot della non-profit Common Crawl, sopra citata); 3) similmente, esistono bot di sviluppatori di IAG la cui finalita' non e' stata esplicitamente dichiarata o di cui non sono stati condivisi i dettagli tecnici, per cui e' difficile conoscere i comportamenti e gli scopi del loro utilizzo (e.g. ClaudeBot di Anthropic).

Conclusione.

L'intelligenza artificiale generativa e' foriera di benefici per la collettivita' che non possono essere limitati, negati, ne' 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 pubblicita', accesso, riuso e di adozione delle misure di sicurezza previste dal RGPD, dovrebbero valutare, caso per caso, quando risulti necessario, in conformita' 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 ne' per metodo, ne' per risultato, possono contenere gli effetti dello scraping finalizzato all'addestramento degli algoritmi di intelligenza artificiale generativa.

(1) Il Garante ha, in passato, dichiarato illecita l'attivita' di web scraping posta in essere dalla societa' 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

(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 autorita' 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.

(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 utilizzera' 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/

(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'autorita' finlandese ha sostenuto l'illiceita' 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

(12) https://www.youtube.com/t/terms#6bedad2de4

(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 e' 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.