GARANTE PER LA PROTEZIONE DEI DATI PERSONALI

SEE ALSO ORDER OF 20 MAY 2024

May 2024

Web scraping and generative artificial intelligence: briefing note and possible enforcement actions

Introduction

With the present document, the Garante intends to provide initial indications on the phenomenon of the massive collection of personal data from the web for the purpose of training generative artificial intelligence models (hereinafter also referred to as 'IAG') and to point out possible counter actions that the managers of websites and online platforms, both public and private, operating in Italy, as data controllers, could implement in order to prevent, where deemed incompatible with the legal bases and purposes of the publication, the collection of data by third parties for the purpose of training artificial intelligence models.

This document only concerns personal data that are disseminated as they are published on websites and online platforms.

The document takes into account the contributions received by the Authority within the framework of the fact-finding investigation into *web scraping*, which was resolved by order of 21 December 2023, published in the Official Gazette No. 14 of 18 January 2024.

In any case, it is up to the operators of the aforementioned public and private websites and platforms, insofar as they are at the same time data controllers of personal data within the meaning of Regulation (EU) 2016/679 (hereinafter 'GDPR'), to make the assessments to be carried out on a case-by-case basis, on the basis of the nature, scope, context and purposes of the personal data processed, the publicity, access and re-use regime to be ensured, the protection afforded by other specific legislation (e.g. copyright protection legislation), taking into account the state of the art (understood in a purely technological sense) and the costs of implementation (in particular with reference to small and medium-sized enterprises).

Web scraping and the right to data protection

To the extent that *web scraping* involves the collection of information traceable to an identified or identifiable natural person, a data protection issue arises.

The focus of *compliance with the* GDPR is generally on entities that process personal data collected through *web scraping* techniques, in particular with regard to the identification of a suitable legal basis under Article 6 of the GDPR for the processing of such data.[1]the identification of which must be carried out on the basis of a suitability assessment that the controller must be able to prove, in accordance with the *accountability* principle set out in Article 5(2) of the GDPR.

This paper proposes a different perspective, examining the position of public and private parties, *website* and *online* platform operators, acting as data controllers, who make publicly available, data (including personal data) that are collected by third-party *bots.*

[1] The Garante has, in the past, declared unlawful the *web scraping* activity carried out by the US company Clearview, [web doc no. 9751362], available at the URL https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9751362 and that carried out by the Trovanumeri platform [web doc no. 9903067], available at the URL https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9903067.

In line with this approach, the document indicates some of the possible precautions that, on the basis of an assessment to be carried out on a case-by-case basis, data controllers may implement in order to prevent or mitigate, in a selective manner, *web scraping* activity for the purpose of training generative artificial intelligence models.

In this regard, it seems appropriate to recall that any personal data controller, whether public or private, under the Regulation may make such personal data available to the public exclusively for specific purposes and on the basis of one or more of the conditions of legitimacy among those laid down in Article 6 of the Regulation (e.g.: transparency obligations, legal publicity, public procedures, right to report, existing contract with data subjects).

The judgement of the lawfulness of *web scraping* must, therefore, be made on a case-by-case basis on the basis of the different and opposing rights at stake: in this sense, for the purposes of this paper, such lawfulness is not and can only be assessed in purely theoretical terms.

It should also be noted that this document does not deal with indicating the security measures that data controllers must implement in order to protect personal data from operations that can be qualified as 'malicious' *web scraping*, insofar as they are able to exploit vulnerabilities in information systems that are not adequately protected from the point of view of IT security. In this respect, the obligation of data controllers under Article 32 of the GDPR to ensure, on a permanent basis, the confidentiality, integrity, availability and resilience of processing systems and services remains firm. In this regard, reference is made to the principles expressed in the decision adopted, in November 2022, by the Irish authority against Meta Platforms Ireland Ltd[2] regarding the failure to adequately protect data (due to non-RPD-compliant settings of the Facebook Search, Facebook Messenger *Contact Importer* and *Instagram Contact Importer* tools) and the subsequent *online* collection, through *web scraping* techniques adopted by third parties, of the data of approximately 533 million users of the Facebook service in the period between 25 May 2018 and September 2019.[3]

Mass data collection techniques from the web and their purposes

The birth and affirmation of the Internet is intrinsically linked to its open technological architecture based on *de facto* computer standards*,* independent of 'proprietary' specifications, founded on the TCP (*Transmission Control Protocol*) and IP (*Internet Protocol*) *suite of* protocols. Over time, these protocols were joined by, among others, the HTTP protocol (*Hyper Text Transfer Protocol*) with which, following the decision by CERN in Geneva to make it public in 1990, the free development of the *World Wide* Web (hereinafter 'web') as well as the

[2] https://www.dataprotection.ie/sites/default/files/uploads/2022-12/Final%20Decision_IN-21-4-2_Redacted.pdf.

[3] The data breach had also been brought to the attention of the public by the Garante through the adoption of a general warning measure addressed to all natural or legal persons, public authorities, services and any body which, individually or jointly with others, performed the role of data controller or processor in the context of the processing of personal data. The order made it clear that any processing of personal data that *was* the subject of the *data breach* at Meta would be in breach of Articles 5(1)(a), 6 and 9 of the Regulation, with all the consequences, including those of a sanctioning nature, envisaged by the rules on the protection of personal data [doc web 9574600]. Available at URL https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9574600.

we know, with the first formalisation in standard form (HTTP/1.1) with the 1997 RFC- 2068 document.

Web surfing is therefore based on open protocols that allow information and data to be found that are publicly available *online* or made available in controlled access areas. Information and data may also be collected in a systematic manner through programmes (*web robots* or, more simply, *bots*) that operate in an automated manner simulating human navigation, provided that the resources (*e.g. websites*, content, etc.) visited by the latter are accessible to the indistinct public and not subject to access controls.

A recent study conducted by Imperva,[4] a company of the French group Thales, revealed that, in the year 2023, 49.6 % of all Internet traffic was generated by *bots*, an increase of 2.1 % compared to the previous year, an increase that was partly attributed to the spread of artificial intelligence systems and, in particular, of the *large language models (*hereinafter also referred to as 'LLM' - *Large Language Models*) underlying generative artificial intelligence.[5]

In the *online* environment, the most well-known *bots* used are the *web crawlers* (also called '*spiders*') of search engines. These are programmes that systematically scan the *web in order to* collect data from *web* pages and index them to ensure the functioning of search engines (GoogleBot and BingBot, for instance, are the search engine *spiders of* Google and Microsoft).

We speak of *web scraping* where the activity of mass and indiscriminate collection of data (including personal data) conducted by means of *web crawling* techniques is combined with an activity consisting of storing and preserving the data collected by the *bots* for subsequent targeted analysis, processing and use.[6]

The purposes for which *bots* are used and *web scraping* activities are carried out are *manifold*, and some are certainly malicious (think of traditional DDoS attacks - *Distributed Denial of Service* - forced *login* attempts, *scalping*, credential theft, and digital fraud), while for these others, the assessment of lawfulness or unlawfulness inevitably remains subject to an assessment to be carried out on a case-by-case basis on the basis of a plurality of evaluations that are the responsibility in some respects of the subject proceeding with the activity and in others of the subject publishing the personal data that are the subject of that activity. Among the purposes underlying the web scraping activity, as mentioned above, there is also that of training generative artificial intelligence algorithms.[7]. The large *datasets* used by generative artificial intelligence developers have varied origins, but *web scraping* constitutes a common denominator. Developers can, in fact, use *datasets that are the* subject of their own *scraping* activities, or draw from third-party *data lakes* (these include, by way of example only, the *open repository* of the US non-profit Common Crawl,[8] the datasets of the French-American platform Hugging Face[9] or of the non-profit

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

[5] To give an idea of the phenomenon, we represent that ten years ago, in 2013, Internet traffic consisted of 23.6% traffic generated by *bad bots (bad bots*), 19.4% by good bots (good bots) and 57% by humans.

[6] For the purposes of this document, the term *web scraping will be* used as including *web crawling*.

[7] Generative artificial intelligence is defined as an artificial intelligence system capable of generating new texts, images, audio and video.

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

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

German LAION AI[10]) which were, in turn, previously created by means of *scraping* operations. On the other hand, it is possible that the training datasets consist of data already held by the developers, such as user data of services offered by the same developer or user data of a social network.

Possible actions against *web scraping for the* purpose of training generative artificial intelligence

Net, therefore, of the obligations currently incumbent on data controllers in connection with both the data disclosure, access and re-use regimes provided for *by law* and the security measures required to ensure data protection, the Garante deems it useful to provide some indications to website and online platform operators, operating in Italy as data controllers of personal data made available to the public through online platforms on the possible precautions that could be adopted to mitigate the effects of third-party *web scraping* aimed at training generative artificial intelligence systems where considered, in implementation of the principle of accountability by the individual data controller, incompatible with the purposes and legal bases of making personal data available to the public.

In the full awareness that none of these measures can be considered suitable to prevent *web scraping* 100%*, they must be considered as* precautions to be adopted on the basis of an autonomous assessment of the data controller, implementing the principle of *accountability*, in order to prevent the deemed unauthorised use by third parties of personal data published as data controller.

1. Creation of restricted areas

Given that the training of generative artificial intelligence is based on enormous quantities of data that often come from direct *web scraping* activities (i.e. carried out by the same subject who develops the model), indirect (i.e. carried out on datasets created through *web scraping* techniques by third parties with respect to the model developer) or hybrid, on sources present on the web, the creation of reserved areas, which can be accessed only after registration, represents a valid caution insofar as it removes data from the considered public availability. This type of technical-organisational caution may, albeit indirectly, contribute to greater protection of personal data than web scraping activities.

On the other hand, such a measure may not give rise to excessive data processing by the controller, in breach of the principle of minimisation set out in Article 5(1)(c) of the GDPR (by way of example, it should be recalled that controllers should not impose additional and unjustified registration burdens on users browsing their *websites* or *online* platforms and using their services [11].

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

[11] In this regard, reference is made to a recent decision, adopted within the framework of the European cooperation procedure *pursuant to* Article 60 ff. of the GDPR, in which the Finnish authority upheld the unlawfulness of the obligation imposed by the data controller to create a user account in order to complete a single online purchase on an e-commerce platform. Available at the 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. Inclusion of *ad hoc* clauses in the terms of service

The inclusion in the Terms of Service (ToS) of a *website* or *online* platform of the express prohibition to use *web scraping* techniques constitutes a contractual clause that, if not respected, allows the operators of said websites and platforms to take legal action to have the contractual breach of contract of the counterparty declared. This is a caution of a purely legal nature that operates, as such, *ex post* facto, but which can act as a special-preventive instrument and, in this way, act as a deterrent, contributing to greater protection of personal data with respect to *web scraping* activities. In this regard, reference is made to the wide use and effectiveness of such a measure, in particular, in the protection of copyright-protected content (mention is made, among many others, of YouTube's terms of service, to which Google prohibits access by automated means, such as robots, botnets or *scraping* tools, unless they are public search engines, in accordance with YouTube's robots.txt file, or unless YouTube has given its prior written consent[12]).

3. Network traffic monitoring

A simple technical precaution such as monitoring HTTP requests received by a website or platform makes it possible to detect any abnormal flow of data in and out of a website or online platform and to take appropriate protective countermeasures. This caution may also be accompanied by *rate limiting*, a technical measure that allows network traffic and the number of requests to be limited by selecting only those coming from certain IP addresses, in order to prevent excessive data traffic (in particular DDoS attacks or *web scraping*) in *advance. These are* technical precautions that, albeit indirectly, can contribute to greater protection of personal data than *web scraping* activities for the purpose of training generative artificial intelligence.

4. Intervention on bots

As illustrated above, *web scraping* is based on the use of bots. Any technique that can limit access to *bots* therefore proves to be an effective method to curb the automated data collection activity that is carried out through such software. It should be emphasised that no technique that acts on *bots* is able to nullify their operation 100%, but also that certain counteracting actions can undoubtedly contribute to preventing/mitigating unwanted *web scraping for the* purpose of training generative artificial intelligence.

In this regard, mention should be made, by way of example only:

i) the inclusion of CAPTCHA (*Completely Automated Public Turing-test-to-tell Computers and Humans Apart*) checks, which, by imposing an action that can only be performed by a human being, prevent bots from operating;

ii) The periodic modification of HTML *markup,* so as to hinder or otherwise make *scraping* by *bots* more complicated. Such modification may be achieved by nesting HTML elements or by modifying other aspects of the *markup*, even in a random manner.

iii) the incorporation of content or data that is to be removed from *scraping* within media objects, such as images (think of the use of

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

this technique in the case of short text such as telephone numbers or *emails*) or other forms of media. In this case, data extraction by the *bot* would be significantly more complex. For instance, for the extraction of data from the image - assuming the *bot was able to* identify its presence there encoded - optical character recognition (OCR) would be required, as the content does not exist as a character string in the code of the *web* page. It should be noted, however, that such a measure, while representing a possible form of subtraction of certain data from the *scraping* activity, could represent an obstacle for users pursuing certain legitimate purposes, (*e.g. the* impossibility of copying content from the *website*).

iv) monitoring of *log files* in order to block unwanted *user-agents,* where identifiable[13];

v) intervention on the robot.txt *file.* The robots.txt *file is a* technical tool that, since June 1994, plays a fundamental role in the management of access to data contained in websites, as it allows managers to indicate whether or not the entire site or certain parts of it may be subject to indexing and *scraping*. Created as a tool to regulate the access of search engine *crawlers* (and thus to control the indexation of websites), the trick based on *robots.txt* (basically, a *black-list* of contents to be removed from indexation) has evolved into the REP (*Robot Exclusion* Protocol), an informal protocol to allow (*allow)* or *disallow (disallow*) access to different types of bots. In the present case, it is theoretically conceivable to include in the robot.txt *file* indications aimed at not allowing (*disallow*) the action of specific *bots* aimed at *scraping for the* purpose of training generative artificial intelligence belonging to certain developers. There are, in fact, certain *bots* which, by self-declaration of the IAG developers themselves, are aimed at *scraping* for such purposes. We mention, by way of example only, the *bots* of OpenAI (GPTBot)[14] and Google's (Google-Extended)[15]which can be excluded, by means of REP, to prevent the total or partial *scraping* of a *website* by its developers. This is a targeted technical measure, but limited in its effectiveness for several reasons, including:1) the REP is not a recognised *standard* and, therefore, its observance is only based on the assumption of an ethical commitment on the part of *web scrapers*; 2) there are *bots* that collect data from the *web* by means of *scraping* techniques for purposes other than exclusively IAG training, and to whose *data lake* IAG developers frequently resort for their own purposes (among these, the best known is certainly the CCBot of the non-profit Common Crawl, mentioned above) 3) Similarly, there are *bots of* IAG developers whose purpose has not been explicitly stated or whose technical details have not been shared, so that it is difficult to know the behaviour and purpose of their use (*e. g. ClaudeBot of Anthropologie, cited above).g.* Anthropic's ClaudeBot).

[13] *User-agents* may also be anonymous or indicate a non-qualifying name or be *spoofed*.

[14] https://platform.openai.com/docs/gptbot.

[15] https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers?hl=it. *Google-Extended* is different from Google's main *crawler* (*Googlebot),* which is used for the operation of Google's search engine, and does not affect the inclusion or *ranking of* a site in that engine.

Conclusion

Generative artificial intelligence is a harbinger of benefits for the community that cannot be limited, denied or diminished. The training of the models underlying the functioning of such systems requires, however, a huge amount of data (even of a personal nature), often originating from a massive and indiscriminate collection carried out on the *web* by means of *web scraping* techniques. Managers of *websites* and *online* platforms that also act as data controllers, without prejudice to the obligations of publicity, access, re-use and adoption of the security measures provided for by the GDPR, should assess, on a case-by-case basis, when it is necessary, in compliance with the current rules, to remove the personal data they process from third-party *bots* by adopting countermeasures such as those indicated which, although not exhaustive in terms of method and result, may contain the effects of *scraping aimed at training generative artificial intelligence algorithms*.