

# Lawfulness of the mass processing of publicly accessible online data to train large language models

Taner Kuru \*

## Key points

- With the unstoppable rise of large language models, the lawfulness of training them with publicly accessible online data started to be questioned. This article examines this question in the European Union data protection context by taking ChatGPT as a case study to understand whether a valid legal basis could be found for these processing activities.
- This discussion has so far focused on Article 6(1)(f) GDPR and inquired whether and to what extent OpenAI has a legitimate interest in processing publicly accessible online data to train ChatGPT. However, this article suggests that Article 9 GDPR should apply to these processing activities considering the relevant judgments of the Court of Justice of the European Union. Accordingly, it examines the exceptions listed in Article 9(2) GDPR that OpenAI could rely on for these processing activities.
- Consequently, the article argues that while it is practically not possible for OpenAI to obtain and demonstrate the explicit consent of data subjects whose publicly accessible personal data are used for training ChatGPT, the pool of personal data that are manifestly made public by data subjects and, therefore, could be used for training large language models appears limited. Considering that OpenAI

or other developers currently do not mention that only this limited amount of personal data is being used for training their models, the article calls for clarification for all stakeholders.

- Lastly, it speculates on a possible way forward and provides an overview of the responsibilities of all stakeholders involved to ensure that these models are developed responsibly.

## Introduction

In early 2020, the New York Times reported that ‘a little-known start-up helps law enforcement match photos of unknown people to their online images’.<sup>1</sup> It was revealed that this start-up, Clearview AI, was operating a facial recognition app that could find the public photos of individuals when a picture of them is uploaded on it, thanks to its database consisting of more than 3 billion images scraped from all over the Internet.<sup>2</sup> Users of this app were reported to be various, ranging from law enforcement authorities to private companies.<sup>3</sup> After this revelation, the company received major criticism and was later scrutinized by judicial and administrative authorities worldwide.<sup>4</sup> Most notably, data protection authorities (DPA) of France, Greece, Italy, and the UK issued Clearview AI with hefty fines of up to €20M.<sup>5</sup>

In these cases, Clearview AI repeatedly relied on the claim that the scraped images were publicly available so

\* Taner Kuru, Tilburg Institute for Law, Technology, and Society (TILT), Tilburg University, Tilburg, The Netherlands. I would like to thank my supervisors Professor Eleni Kosta and Professor Paul de Hert for their valuable comments on an earlier version of this article.

1 Kashmir Hill, 'The Secretive Company That Might End Privacy as We Know It' *The New York Times* (New York, 18 January 2020) <<https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>> accessed 31 October 2023.

2 Ibid.

3 Ibid.

4 Isadora N Rezende, 'Facial Recognition in Police Hands: Assessing the  
'Clearview Case' from a European Perspective' (2020) 11 New Journal of  
European Criminal Law 375, 376; Gaurav Pathak, 'Manifestly Made  
Public: Clearview and GDPR' (2022) 8 European Data Protection Law  
Review 419, 419.

5 ‘Facial Recognition: 20 Million Euros Penalty against CLEARVIEW AI’ (CNIL, 20 October 2022) <<https://www.cnil.fr/en/facial-recognition-20-million-euros-penalty-against-clearview-ai>> accessed 31 October 2023; ‘Facial Recognition: Italian SA Fines Clearview AI EUR 20 Million’ (European Data Protection Board, 10 March 2022) <[https://edpb.europa.eu/news/national-news/2022/facial-recognition-italian-sa-fines-clear-view-ai-eur-20-million\\_en](https://edpb.europa.eu/news/national-news/2022/facial-recognition-italian-sa-fines-clear-view-ai-eur-20-million_en)> accessed 31 October 2023; ‘Hellenic DPA Fines Clearview AI 20 Million Euros’ (European Data Protection Board, 20 July 2022) <[https://edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros\\_en](https://edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros_en)> accessed 31 October 2023; ‘ICO Fines Facial Recognition Database Company Clearview AI Inc More than £7.5m and Orders UK Data to be Deleted’, (Information Commissioner’s Office, 23 May 2022) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-data-base-company-clearview-ai-inc/>> accessed 31 October 2023.

that no privacy interest of the individuals concerned could be claimed.<sup>6</sup> However, such an argument did not save them from being severely fined in these proceedings as, to the best of our knowledge, only the UK data protection watchdog's decision was overturned but for jurisdictional reasons.<sup>7</sup> Unsurprisingly, after the revelation of Clearview AI, the debate on the processing of publicly accessible online data became a hot topic in the data protection context, especially given the controversial cases and emerging business models built on processing these data.<sup>8</sup> And with the unstoppable rise of large language models (LLMs), it has already become the next big thing.

Since its launch in late 2022, OpenAI's ChatGPT has been 'the thing' that everyone talks about, proven by the fact that it became the fastest-growing consumer application in history by reaching 100 million monthly active users two months after its launch.<sup>9</sup> Simply put, LLMs like ChatGPT are trained on billions of words available in various (scraped) sources, such as news articles, books, personal blogs, or derived from user interaction, and tasked with recognizing text and generating human-like responses to queries as they are enabled to

predict the following words through their learning progress. Accordingly, they are being used for various purposes. Furthermore, some of these models, especially ChatGPT, started offering their users additional functions, such as image processing and generation, thanks to their integrated tools.<sup>10</sup> Like any other technological advancement, however, LLMs, specifically ChatGPT, are not free from controversy. For example, whether these models could replace some professions, how teachers could detect student essays written by these models, how to prevent these models from producing misleading or defamatory answers, how to mitigate the risk of bias and adverse impact on the environment, and whether and to what extent private parties should be allowed to capitalize on the information that Internet users have commonly created are some of the questions that are posed to underline the risks attached to these models.<sup>11</sup>

The most pressing legal concerns, however, are attached to the fact that a vast amount of (online) data is used for training these models, especially ChatGPT. Although OpenAI is secretive about the training datasets of ChatGPT,<sup>12</sup> there are several clues about what

- 6 Pathak (n 4) 420; Brendan Walker-Munro, 'Hyper-Collection: A Possible New Paradigm in Modern Surveillance' (2023) 21 *Surveillance & Society* 120, 128; Geoffrey Xiao, 'Bad Bots: Regulating the Scraping of Public Personal Information' (2021) 34 *Harvard Journal of Law & Technology* 701, 702.
- 7 Natasha Lomas, 'Selfie-scraper, Clearview AI, Wins Appeal against UK Privacy Sanction' (*TechCrunch*, 18 October 2023) <<https://techcrunch.com/2023/10/18/clearview-wins-ico-appeal/>> accessed 31 October 2023. Note that the Information Commissioner's Office announced that it was seeking permission to appeal this ruling, see, 'Information Commissioner Seeks Permission to Appeal Clearview AI Inc Ruling' (*Information Commissioner's Office*, 17 November 2023) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/11/information-commissioner-seeks-permission-to-appeal-clearview-ai-inc-ruling/>> accessed 31 May 2024. It should further be noted here that Clearview AI has also settled a class action lawsuit against it in the United States, see, Sara Merken, 'Clearview AI Strikes 'Unique' Deal to End Privacy Class Action' *Reuters* (London, 13 June 2024) <<https://www.reuters.com/legal/litigation/clearview-ai-strikes-unique-deal-end-privacy-class-action-2024-06-13/>> accessed 5 August 2024.
- 8 See eg, Kashmir Hill, 'A Face Search Engine Anyone Can Use Is Alarming Accurate' *The New York Times* (New York, 26 May 2022) <<https://www.nytimes.com/2022/05/26/technology/pimeyes-facial-recognition-search.html>> accessed 28 May 2024; Derk Stokmans and Stefan Vermeulen, 'Belastingdienst verzamelde op grote schaal persoonlijke gegevens, ook op sociale media' (*NRC*, 13 September 2023) <<https://www.nrc.nl/nieuws/2023/09/13/extern-onderzoek-naar-omstreden-database-belastingdienst-a4174396>> accessed 31 October 2023.
- 9 Krystal Hu, 'ChatGPT Sets Record for Fastest-growing User Base—analyst Note' *Reuters* (London, 2 February 2023) <<https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>> accessed 31 October 2023.
- 10 'ChatGPT Capabilities Overview' (*OpenAI*) <<https://help.openai.com/en/articles/9260256-chatgpt-capabilities-overview>> accessed 28 May 2024.
- 11 See, Pranshu Verma and Gerrit De Vynck, 'ChatGPT Took Their Jobs. Now they Walk Dogs and Fix Air Conditioners' *The Washington Post* (Washington DC, 2 June 2023) <<https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/>> accessed 31 October 2023; Jocelyn Gecker and The Associated Press, 'College Professors are in 'Full-on Crisis mode' as they Catch One 'ChatGPT Plagiarist' after Another' (*Fortune*, 10 August 2023) <<https://fortune.com/2023/08/10/chatgpt-cheating-plagiarism-college-professors-full-on-crisis-mode/>> accessed 31 October 2023; Byron Kaye 'Australian Mayor Readies World's First Defamation Lawsuit over ChatGPT Content' *Reuters* (London, 5 April 2023) <<https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/>> accessed 31 October 2023; James Vincent, 'OpenAI Sued for Defamation after ChatGPT Fabricates Legal Accusations against Radio Host' (*The Verge*, 9 June 2023) <<https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit>> accessed 31 October 2023; Matthias C Rillig and others, 'Risks and Benefits of Large Language Models for the Environment' (2023) 57 *Environmental Science & Technology* 3464; Norwegian Consumer Council, 'Ghost in the Machine—Addressing the Consumer Harms of Generative AI' (2023) 17–37; Will D Heaven, 'These Six Questions will Dictate the Future of Generative AI' (*MIT Technology Review*, 19 December 2023) <<https://www.technologyreview.com/2023/12/19/1084505/generative-ai-artificial-intelligence-bias-jobs-copyright-misinformation/>> accessed 31 May 2024.
- 12 Clothilde Goujard and Gian Volpicelli, 'ChatGPT is Entering a World of Regulatory Pain in Europe' *Politico* (Arlington County, VA, 10 April 2023) <<https://www.politico.eu/article/chatgpt-world-regulatory-pain-eu-privacy-data-protection-gdpr/>> accessed 31 October 2023; Jess Weatherbed, 'OpenAI's Regulatory Troubles are Only Just Beginning' (*The Verge*, 5 May 2023) <<https://www.theverge.com/2023/5/5/23709833/openai-chatgpt-gdpr-ai-regulation-europe-eu-italy>> accessed 31 October 2023; Zachary Small, 'Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement' *The New York Times* (New York, 10 July 2023) <<https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html>> accessed 31 October 2023; Kevin Schaul, Szu Y Chen and Nitasha Tiku, 'Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart' *The Washington Post* (Washington DC, 19 April 2023) <<https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>> accessed 31 October 2023; Matt Burges, 'ChatGPT has a Big Privacy Problem' (*WIRED*, 4 April 2023) <<https://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr/>> accessed 31 October 2023; Norwegian Consumer Council (n 11) 47.

kind of data and sources have been used for its training. For example, OpenAI states that its ‘language models are trained on a broad corpus of text that includes publicly available content, licensed content, and content generated by human reviewers’ and ‘(...) some of our training data includes personal information that is available on the public internet’.<sup>13</sup> Combining these statements with the existence of the GPTBot,<sup>14</sup> OpenAI’s web crawler, it can be inferred that ChatGPT is trained on, among others, data scraped from various publicly accessible websites. In other words, OpenAI processes massive amounts of online data that is publicly accessible to train ChatGPT. It has been reported that several other developers have also followed the same approach to train their LLMs.<sup>15</sup> Furthermore, it was recently announced that ChatGPT can now browse the Internet to provide its answers.<sup>16</sup> Accordingly, it is claimed that ‘If you’ve posted anything even remotely personal in English on the internet, chances are your data might be part of some of the world’s most popular LLMs.’<sup>17</sup>

Unsurprisingly, these practices did not take long to be legally challenged. In the USA, several lawsuits were filed against OpenAI in which the plaintiffs claimed, among others, that ChatGPT was trained on their copyrighted work without permission.<sup>18</sup> As a

result of this tension, it is observed that more and more websites, especially publishers, are blocking ChatGPT’s web crawler today.<sup>19</sup> On the other side of the Atlantic, like in the Clearview AI case, DPAs are taking the lead in scrutinizing these practices. For example, the Italian DPA made its way to the international headlines in early 2023 when it blocked OpenAI from processing Italian users’ data due to several violations of the General Data Protection Regulation (GDPR),<sup>20</sup> including the lack of ‘legal basis underpinning the massive collection and processing of personal data in order to “train” the algorithms on which the platform relies’.<sup>21</sup> Although OpenAI has complied with the Italian DPA’s requests to lift this ban and become operational again, this saga appears to have restarted as the Italian DPA recently announced that it has notified several breaches of the GDPR to OpenAI and launched an investigation into its new text-to-video tool, Sora, in which they asked OpenAI to clarify, among others, the legal basis for processing (sensitive) personal data to train its models.<sup>22</sup> In the meantime, while several other DPAs have announced their investigations of ChatGPT, the European Data Protection Board (EDPB) has also created a special task force on ChatGPT to foster

13 ‘Our Approach to AI Safety’ (OpenAI, 5 April 2023) <<https://openai.com/blog/our-approach-to-ai-safety>> accessed 31 October 2023.

14 ‘GPTBot’ (OpenAI) <<https://platform.openai.com/docs/gptbot>> accessed 31 October 2023.

15 Schaul, Chen and Tiku (n 12).

16 Antoinette Radford and Zoe Kleinman, ‘ChatGPT Can Now Access Up to Date Information’ (BBC, 27 September 2023) <<https://www.bbc.com/news/technology-66940771>> accessed 31 October 2023. Although the browsing feature is currently not available for all users, it is expected that OpenAI will roll out it for all users soon. It should also be noted that this feature had already been implemented; however, the users benefited from this feature to bypass paywalls through this feature, which caused OpenAI to pull it back, see Wes Davis, ‘ChatGPT Can Now Search the Web in Real Time’ (The Verge, 27 September 2023) <<https://www.theverge.com/2023/9/27/23892781/openai-chatgpt-live-web-results-browse-with-bing>> last accessed on 31 October 2023.

17 Melissa Heikkilä, ‘What does GPT-3 “Know” about Me?’ (MIT Technology Review, 31 August 2022) <<https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/>> accessed 12 October 2023. See further, Matt Burgess and Reece Rogers, ‘How to Stop Your Data from Being Used to Train AI’ (WIRED, 10 April 2024) <<https://www.wired.com/story/how-to-stop-your-data-from-being-used-to-train-ai/>> accessed 31 May 2024.

18 Small (n 12); Tom Gerken and Liv McMahon, ‘Game of Thrones Author Sues ChatGPT Owner OpenAI’ (BBC, 21 September 2023) <<https://www.bbc.com/news/technology-66866577>> accessed 31 October 2023. It has been reported that similar lawsuits could follow soon, see Irina Ivanova, ‘What if OpenAI Trained ChatGPT with Illegal Data Scraping? The New York Times is Reportedly Considering Suing to Put that to the Test’ (Fortune, 17 August 2023) <<https://fortune.com/2023/08/17/openai-new-york-times-lawsuit-illegal-scraping/>> accessed 31 October 2023. Recently, a similar lawsuit was also filed against Google Bard, see Gianluca Campus, ‘Generative AI: the US Class Action against Google Bard (and other AI tools) for Web Scraping’ (Kluwer Copyright Blog, 3

October 2023) <<https://copyrightblog.kluweriplaw.com/2023/10/03/generative-ai-the-us-class-action-against-google-bard-and-other-ai-tools-for-web-scraping/>> accessed 31 October 2023.

19 Dan Milmo ‘The Guardian Blocks ChatGPT Owner OpenAI from Trawling its Content’ *The Guardian* (London, 1 September 2023) <<https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content>> accessed 31 October 2023; Julia Tar, ‘Several French Media Block OpenAI’s GPTBot Over Data Collection Concerns’ (EURACTIV, 29 August 2023) <<https://www.euractiv.com/section/artificial-intelligence/news/several-french-media-block-openais-gptbot-over-data-collection-concerns/>> accessed 31 October 2023; Lottie Hayton, ‘BBC Blocks ChatGPT from Using its Content’ *The Times* (London, 7 October 2023) <<https://www.thetimes.co.uk/article/bbc-blocks-chatgpt-from-using-its-content-358wz5kvj>> accessed 31 October 2023.

20 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ 2016 L 119/1.

21 ‘Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell’età dei minori’ (Garante per la protezione dei dati personali, 31 March 2023) <<https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847?english>> accessed 28 May 2024.

22 ChatGPT: Garante privacy, notificato a OpenAI l’atto di contestazione per le violazioni alla normativa privacy’ (Garante per la protezione dei dati personali, 29 January 2024) <<https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9978020>>; ‘Intelligenza artificiale, il Garante privacy avvia istruttoria su “Sora” di OpenAI. Chieste alla società informazioni su algoritmo che crea brevi video da poche righe di testo’ (Garante per la protezione dei dati personali, 8 March 2024) <<https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9991867>> accessed 28 May 2024.

cooperation and exchange information on possible enforcement actions conducted to that end.<sup>23</sup>

At the climax of the legal battles on the mass processing of publicly accessible online data for training LLMs, this article focuses on the most fundamental question in the European Union (EU) data protection context by taking ChatGPT as a case study<sup>24</sup>: whether a valid legal basis could be found for these processing practices. To that end, it first explains that OpenAI's mass processing of publicly accessible online data to train ChatGPT falls under the scope of the EU data protection framework and, therefore, there needs to be a valid legal basis for these processing activities. Accordingly, it examines the current discussion on OpenAI's reliance on legitimate interest, enshrined in Article 6(1)(f) GDPR, as the legal basis for these processing activities. It then demonstrates why Article 9 GDPR should be evaluated to find the proper legal basis for these processing activities instead of Article 6 GDPR in light of the recent case law of the Court of Justice of the European Union (CJEU) since OpenAI processes sensitive and non-sensitive data *en bloc* when training ChatGPT with publicly accessible online data. The article then examines which exceptions enshrined in Article 9(2) GDPR OpenAI could rely on for training ChatGPT with this method. It first explains why it is practically impossible for OpenAI to obtain and demonstrate the explicit consent of data subjects whose personal data is scraped from the Internet by OpenAI to train ChatGPT. As a second option, the article explores the 'manifestly made public by the data subject' exception; however, it reaches the conclusion that the pool of personal data that are manifestly made public by data subjects and, therefore, could be used for training LLMs appears limited. Given that there has been no explicit reference made by OpenAI or any other developer that only this limited amount of personal data is being used for training their LLMs, the article calls for clarification on this matter to provide clarity

for all stakeholders. Lastly, the article speculates on a possible way forward and concludes by underlining the responsibilities of all stakeholders involved to ensure that LLMs are developed without undermining the right to data protection of individuals concerned even when, if it exists, a valid legal basis is found.

## Personal data processing by OpenAI when training ChatGPT with publicly accessible online data

OpenAI states that ‘OpenAI’s large language models, including the models that power ChatGPT, are developed using three primary sources of information: (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or our human trainers provide.’<sup>25</sup> The relevance of this statement for the EU data protection framework is that all these sources contain personal data, and processing them triggers the application of the GDPR. However, since the focus of this article is the first source mentioned by OpenAI, namely the ‘information that is publicly available on the internet’, personal data processing activities conducted by OpenAI are demonstrated from this perspective.

Personal data is defined as ‘any information relating to an identified or identifiable natural person’ in the GDPR.<sup>26</sup> Accordingly, various kinds of information may fall under the scope of this definition, such as names, e-mail addresses, license plates, usernames, IP addresses, and so on. Furthermore, in its jurisprudence, the CJEU has provided an expansive interpretation of the definition of personal data, causing a significant expansion in the GDPR’s material scope.<sup>27</sup> Based on this case law, for example, dynamic IP addresses and an examiner’s comments on an individual’s answers may constitute personal data.<sup>28</sup>

23 ‘EDPB Resolves Dispute on Transfers by Meta and Creates Task Force on ChatGPT’ (*European Data Protection Board*, 13 April 2023) <[https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt\\_en](https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en)> accessed 1 November 2023; ‘Artificial Intelligence: The Action Plan of the CNIL’ (CNIL, 16 May 2023) <<https://www.cnil.fr/en/artificial-intelligence-action-plan-cnil>> accessed 1 November 2023; Marc Rees, ‘Premières Plaintes Françaises Contre ChatGPT’ (*L’Informé*, 4 May 2023) <[https://www.linforme.com/tech-telecom/article/premieres-plaintes-francaises-contre-chatgpt\\_538.html](https://www.linforme.com/tech-telecom/article/premieres-plaintes-francaises-contre-chatgpt_538.html)> accessed 1 November 2023; Natasha Lomas, ‘Spain’s Privacy Watchdog Says It’s Probing ChatGPT Too’ (*TechCrunch*, 13 April 2023) <<https://techcrunch.com/2023/04/13/chatgpt-spain-gdpr/>> accessed 1 November 2023; ‘Germany Launches Data Protection Inquiry Over ChatGPT’ (*Barron’s*, 24 April 2023) <<https://www.barrons.com/news/germany-launches-data-protection-inquiry-over-chatgpt-cdl5588r>> accessed 1 November 2023; ‘AP Vraagt om Opheldering Over ChatGPT’ (*Autoriteit Persoonsgegevens*, 7 June 2023) <<https://www.autoriteitpersoonsgegevens.nl/actueel/ap-vraagt-om-opheldering-over-chatgpt>> accessed 1 November 2023; Natasha Lomas, ‘Poland Opens Privacy

Probe of ChatGPT following GDPR Complaint' (*TechCrunch*, 21 September 2023) <<https://techcrunch.com/2023/09/21/poland-chatgpt-gdpr-complaint-probe/>> accessed 1 November 2023.

24 ChatGPT was chosen as the case study for this research as it was the most commonly used LLM available to the public at the time of writing this article (August–October 2023).

25 ‘How ChatGPT and Our Language Models are Developed’ (*OpenAI*)  
 <<https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed>> accessed 28 May 2024.

26 GDPR, art 4(1).

27 See, J.M., Case-579/21, [2023] ECLI:EU:C:2023:501, paras 42–45. See further, Nadezhda Purtova, ‘The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law.’ (2018) 10 Law, Innovation and Technology 40.

28 See, *Peter Nowak v Data Protection Commissioner*, Case C-434/16, [2017] ECLI:EU:C:2017:994; *Patrick Breyer v Bundesrepublik Deutschland*, Case C-582/14, [2016] ECLI:EU:C:2016:779.

When training ChatGPT with the ‘information that is publicly available on the internet’, OpenAI inevitably processes data that is considered personal data as per Article 4(1) GDPR.<sup>29</sup> OpenAI itself confirms this by stating, ‘A large amount of data on the internet relates to people, so our training information does incidentally include personal information.’<sup>30</sup> In fact, when researchers managed to expose some of ChatGPT’s training data, they revealed, among others, phone and fax numbers, e-mail and physical addresses, social media handles, names, birthdays, and even some explicit content coming from different corners of the Internet.<sup>31</sup> Regarding the personal data contained in the training datasets of ChatGPT, OpenAI states that

“Web pages crawled with the GPTBot user agent (...) are filtered to remove sources that (...) are known to primarily aggregate personally identifiable information (PII).”<sup>32</sup>

However, it is also acknowledged that “(...) some of our training data includes personal information that is available on the public internet (...) So we work to remove personal information from the training dataset where feasible, fine-tune models to reject requests for personal information of private individuals, and respond to requests from individuals to delete their personal information from our systems. These steps minimize the possibility that our models might generate responses that include the personal information of private individuals.”<sup>33</sup>

The phrases in this statement such as ‘where feasible’ and ‘minimize the possibility’ only support the conclusion that ChatGPT is being trained on and could generate personal data. Furthermore, OpenAI mentions that ‘Our models may learn from personal information to (...) learn about famous people and public figures. This makes our models better at providing relevant responses’,<sup>34</sup> which only enhances this conclusion. Therefore, it is safe to say that OpenAI processes personal data when training ChatGPT with ‘information that is publicly available on the internet’.<sup>35</sup>

Regarding the training phase, OpenAI states that

ChatGPT does not copy or store training information in a database. Instead, it learns about associations between words, and those learnings help the model update its numbers/weights. The model then uses those weights to predict and generate new words in response to a user request. It does not ‘copy and paste’ training information—much like a person who has read a book and sets it down, our models do not have access to training information after they have learned from it.<sup>36</sup>

Even though OpenAI claims with this statement that it does not copy or store the data used for training purposes, this does not change the fact that their operations conducted on ‘information that is publicly available on the internet’ to train ChatGPT still count as processing as defined in the GDPR. This is because Article 4(2) GDPR defines processing as

any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

As can be inferred from the use of ‘any operation’ and ‘such as’ in this provision, the European legislator provides a broad definition of this concept.<sup>37</sup> Given this broad definition, it is safe to conclude that operations carried out by OpenAI to make ChatGPT learn about associations between words constitute processing as per Article 4(2) GDPR, even though the training data are not copied or stored in a database. At this point, it should be mentioned that the CJEU has already stated that the collection of personal data from documents in the public domain as well as the activities of a search engine in exploring the Internet automatically, constantly, and systematically in search of the information

29 See further, European Data Protection Board, ‘Report of the Work Undertaken by the ChatGPT Taskforce’ (23 May 2024) para 15, page 6.

30 OpenAI, ChatGPT (n 25)

31 Milad Nasr and others, ‘Scalable Extraction of Training Data from (Production) Language Models’ (2023) arXiv preprint arXiv:2311.17035 <<https://arxiv.org/pdf/2311.17035.pdf>>

32 OpenAI, GPTBot (n 14)

33 OpenAI, AI safety (n 13)

34 OpenAI, ChatGPT (n 25). OpenAI also confirms, ‘As a result of learning language, ChatGPT responses may sometimes include personal information about individuals whose personal information appears multiple times on the public internet (for example, public figures).’ See, OpenAI, ChatGPT (n 25).

35 The same conclusion could also apply to other LLMs developed by other actors than OpenAI as they also train their models with the same methods. See eg, Mike Clark, ‘Privacy Matters: Meta’s Generative AI Features’

(Meta, 27 September 2023) <<https://about.fb.com/news/2023/09/privacy-matters-metas-generative-ai-features/>> accessed 1 November 2023. See also, ‘Since it takes such a large amount of data to teach effective models, a combination of sources are used for training. These sources include information that is publicly available online and licensed information, as well as information from Meta’s products and services. (...) When we collect public information from the internet or license data from other providers to train our models, it may include personal information. For example, if we collect a public blog post it may include the author’s name and contact information.’ See, ‘How Meta Uses Information for Generative AI Models. What is Generative AI?’ (Meta) <<https://www.facebook.com/privacy/genai>> accessed 1 November 2023.

36 OpenAI, ChatGPT (n 25)

37 See, C-579/21, J.M. (n 27) para 46; *RK v Ministerstvo zdravotníctví*, Case C-659/22, [2023] ECLI:EU:C:2023:745, paras 27–28; *Endemol Shine Finland Oy*, Case 740/22, [2024] ECLI:EU:C:2024:216, para 29.

published therein constitute processing.<sup>38</sup> The similarities of the operations conducted by OpenAI on the ‘information that is publicly available on the internet’ to train ChatGPT with the activities considered as processing by the CJEU in these cases only enhance this conclusion.

## Can OpenAI rely on Article 6(1)(f) GDPR to process publicly accessible online data to train ChatGPT?

OpenAI has a European establishment,<sup>39</sup> provides services to data subjects in the EU,<sup>40</sup> and does not fall under the scope of any exemptions provided for the material scope of the GDPR. Therefore, when processing personal data, it must comply with several requirements outlined in the GDPR. One of the most important obligations in that regard is complying with the lawfulness principle enshrined in Article 5(1)(a) GDPR, which requires processing activities to rely on a valid legal basis. Accordingly, OpenAI should rely on one of the six legal bases listed in Article 6(1) GDPR for training ChatGPT with the ‘information that is publicly available on the internet’.<sup>41</sup>

Whether OpenAI complies with this obligation has already been subject to legal scrutiny. As mentioned above, the Italian DPA blocked OpenAI in early 2023 from processing the personal data of people in Italy due to, among others, the lack of a valid legal basis for processing personal data for training ChatGPT.<sup>42</sup> While doing so, the Italian DPA required OpenAI to change the legal basis for processing personal data for training purposes by removing any reference to the contract, corresponding to the legal basis enshrined in Article 6(1)(b) GDPR that allows personal data to be processed where it is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a

contract, and relying on consent or legitimate interest instead.<sup>43</sup> OpenAI complied with this request and adjusted its privacy policy, which stated, when addressing users from the European Economic Area, Switzerland, and the UK, that ‘Our legal bases for processing your Personal Information include: (...) Our legitimate interests in (...) developing, improving, or promoting our Services, including when we train our models.’<sup>44</sup> Although it is observed that OpenAI updated this privacy policy on 15 December 2023, which became effective as of 15 February 2024, the updated version similarly refers to the legitimate interests of OpenAI, third parties, and broader society as the legal basis for processing several types of personal data including the ‘Data [OpenAI] Receive From Other Sources’, which contain ‘information that is publicly available on the internet’, to train OpenAI’s models.<sup>45</sup>

However, these statements only clarify the legal basis for processing the personal data of ChatGPT users. Given that OpenAI processes ‘information that is publicly available on the internet’, it is clear that the personal data of individuals who are not ChatGPT users are also subjected to this processing. Therefore, OpenAI should also have a valid legal basis for processing the personal data of these individuals. An article published on OpenAI’s website sheds light on this issue by stating,

(...) We use training information lawfully. (...) and the primary sources of this training information are already publicly available. For these reasons, we base our collection and use of personal information that is included in training information on legitimate interests under privacy laws like the GDPR.<sup>46</sup>

This demonstrates that OpenAI relies again on Article 6(1)(f) GDPR as the legal basis for processing the personal data of these individuals.<sup>47</sup>

38 See, *Tietosuojavaltuutettu v Satakunnan Markkinapörssi Oy and Satamedia Oy*, Case C-73/07, [2008] ECLI:EU:C:2008:727, para 37; *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD)*, *Mario Costeja González*, Case C-131/12, [2014] ECLI:EU:C:2014:317, para 28.

39 ‘Introducing OpenAI Dublin’ (OpenAI, 13 September 2023) <<https://openai.com/index/introducing-openai-dublin/>> accessed 28 May 2024. See, GDPR, art 3(1). See further, Scott Ikeda, ‘OpenAI Shifts EU Data Privacy Responsibility to Dublin Office’ *CPO Magazine* (Singapore, 10 January 2024) <<https://www.cpomagazine.com/data-protection/openai-shifts-eu-data-privacy-responsibility-to-dublin-office/>> accessed 28 May 2024.

40 See, GDPR, art 3(2)(a).

41 At this point, it should be noted that OpenAI’s processing activities conducted on the ‘information that is publicly available on the internet’ to train ChatGPT cannot be considered as compatible with the initial purposes for which these data have been processed as per art 6(4) GDPR, due to the lack of any link between OpenAI’s purposes to process these data and the initial purposes, the absence of any relationship between OpenAI and the data subjects concerned, the fact that several types of

personal data are being subjected to these processing activities, and the possible adverse effects of these processing activities on the fundamental rights and freedoms of the data subjects. This conclusion prevents OpenAI from relying on the initial legal bases used by third parties to process these data, meaning that OpenAI should have a (new) legal basis to process these data for training ChatGPT.

42 Garante (n 21)

43 ‘Provvedimento dell’11 aprile 2023 [9874702]’ (*Garante per la protezione dei dati personali*, 11 April 2023) <<https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9874702>> accessed 1 November 2023.

44 ‘Privacy Policy’ (OpenAI, 23 June 2023) <<https://openai.com/policies/jun-2023-privacy-policy/>> accessed 28 May 2024.

45 ‘Europe Privacy Policy’ (OpenAI, 15 December 2023) <<https://openai.com/policies/eu-privacy-policy/>> accessed 28 May 2024.

46 OpenAI, ChatGPT (n 25).

47 Meta also states:

We are committed to being transparent about the legal bases that we use for processing information. We believe use of this information is

Enshrined in Article 6(1)(f) GDPR, this legal basis allows personal data to be processed when the processing is necessary for a legitimate interest of the data controller or a third party unless the interest or fundamental rights and freedoms of the data subjects concerned override such an interest. As can be inferred from this definition, it is not enough for data controllers to claim to have a legitimate interest to rely on this legal basis; they need to pass a so-called three-step test in which they need to cumulatively demonstrate that (i) the purpose of their processing is legitimate, (ii) the processing is necessary for that purpose, and (iii) the interests and the fundamental rights and freedoms of the data subjects concerned are not overriding their interests claimed to be legitimate.<sup>48</sup> Accordingly, data controllers relying on this legal basis, including OpenAI, should demonstrate that they pass this three-step test to be able to process personal data in compliance with the lawfulness principle enshrined in Article 5(1)(a) GDPR.

While the reliance on this legal basis for processing publicly accessible online data, including for the purposes of training AI models, has already been discussed in case law,<sup>49</sup> DPA guidelines,<sup>50</sup> and literature,<sup>51</sup> whether OpenAI could rely on Article 6(1)(f) GDPR to train ChatGPT with ‘information that is publicly available on the internet’ has also started to be questioned.<sup>52</sup> In order to answer this question, the three-step test should be applied to OpenAI’s processing activities conducted on these data to train ChatGPT, and the CJEU’s landmark *Meta v Bundeskartellamt*<sup>53</sup> judgment should be used as guidance in such an effort.

Regarding the first step, it should be underlined that the CJEU has already acknowledged in its case law that, in principle, a wide range of interests are entitled to be

considered legitimate,<sup>54</sup> while it should also be noted that there is currently a pending case before it inquiring whether any interest, especially purely commercial interests, could be considered as such.<sup>55</sup> Nevertheless, in the *Meta v Bundeskartellamt* case, the CJEU examined, among others, whether ‘product improvement’ could constitute a legitimate interest for a data controller. It acknowledged that a data controller’s interest in improving its product or service to make it more efficient and, therefore, attracting consumers could be considered a legitimate interest in processing personal data.<sup>56</sup> Accordingly, OpenAI could claim that it has a legitimate interest in processing the publicly accessible online data to train and further develop ChatGPT, which would allow it to pass the first step. Besides, it is also possible for OpenAI to pass this first step by relying on a third party’s legitimate interest, such as society’s interest in accessing the information, as its updated privacy policy also implies.<sup>57</sup>

Regarding the second step, the CJEU requires an assessment of whether the legitimate interest pursued cannot be reasonably achieved as effectively by other means that interfere less with the fundamental rights and freedoms of data subjects, which should be done by specifically checking whether the data minimization principle, enshrined in Article 5(1)(c) GDPR, is respected.<sup>58</sup> In the *Meta v Bundeskartellamt* case, the processing activities in question were conducted by Meta to improve their products and services by, in a nutshell, creating profiles of data subjects through processing their personal data, including those available in third-party websites and apps.<sup>59</sup> Here, the CJEU underlined the massive scale of this processing activity and its significant impacts on data subjects; hence, it questioned whether

in the legitimate interests of Meta, our users, and other people. In the European region and the United Kingdom, we rely on the basis of legitimate interests to collect and process any personal information included in the publicly available and licensed sources, as well as information people have shared on Meta’s Products and services, to develop and improve AI at Meta.

See, ‘How Meta Uses Information for Generative AI models. What is Generative AI?’ (*Meta*) <<https://www.facebook.com/privacy/genai>> accessed 29 May 2024.

48 *Meta Platforms Inc., Meta Platforms Ireland Ltd, Facebook Deutschland GmbH v Bundeskartellamt*, Case C-252/21, [2023] ECLI:EU:C:2023:537, para 106; *UF, AB v Land Hessen*, Cases C-26/22 and C-64/22, [2023] ECLI:EU:C:2023:958, para 75.

49 C-131/12, *Google Spain SL and Google Inc.* (n 38)

50 ‘Generative AI First Call for Evidence: The Lawful Basis for Web Scraping to Train Generative AI Models’ (*Information Commissioner’s Office*) <<https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/>> accessed 29 May 2024; ‘An organisation builds a training dataset by collecting comments made public and freely accessible by online users on forums, blogs and websites. The purpose of this processing is to design an AI system to evaluate and predict the appreciation of works of art by the general public. In this case, the interest of the organisation in developing and possibly

marketing an AI system may be considered legitimate.’ See, ‘Ensuring the Lawfulness of the Data Processing’ (CNIL, 16 October 2023) <<https://www.cnil.fr/en/ensuring-lawfulness-data-processing>> accessed 1 November 2023.

51 See, for instance, Pablo T Kramcsák, ‘Can Legitimate Interest be an Appropriate Lawful Basis for Processing Artificial Intelligence Training Datasets?’ (2023) 48 *Computer Law & Security Review* 105765.

52 Norwegian Consumer Council (n 11) 47; EDPB, ChatGPT (n 29) paras 16–17, pages 6–7.

53 C-252/21, *Meta* (n 48).

54 C-26/22 and C-64/22, *UF, AB* (n 48), para 76.

55 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens*, C-621/22 [2022].

56 C-252/21, *Meta* (n 48) para 122.

57 See also, C-26/22 and C-64/22, *UF, AB* (n 48), paras 82–83. See further, ‘Developers who rely on broad societal interests need to ensure that those interests are actually being realised rather than assumed, by applying appropriate controls and monitoring measures on the use of the generative AI models they build on web-scraped data.’ ICO, Call for Evidence (n 50).

58 C-252/21, *Meta* (n 48) paras 108–09; See also, C-26/22 and C-64/22, *UF, AB* (n 48), paras 77–78; EDPB, ChatGPT (n 29) para 16, page 6.

59 C-252/21, *Meta* (n 48) paras 26–28.

such processing could be considered necessary for 'product improvement' purposes.<sup>60</sup> This consideration, however, has serious implications for OpenAI. It is self-evident that OpenAI's mass processing of publicly accessible online data to train ChatGPT cannot be considered compliant with the data minimization principle, which requires the personal data to be adequate, relevant, and limited to what is necessary for the purposes of the processing activity in question, given the massive scale and indiscriminate nature, regarding the type of (personal) data that being processed, of this processing.

Nevertheless, it is also observed that, for example, the Information Commissioner's Office (ICO), the UK's data protection watchdog, stated that

currently, most generative AI training is only possible using the volume of data obtained through large-scale scraping. Even though future technological developments may provide novel solutions and alternatives, currently there is little evidence that generative AI could be developed with smaller, proprietary databases.<sup>61</sup>

Such an acknowledgement could lead OpenAI to pass the second step. However, it should be underlined here that although the developers of these models similarly claimed that it is impossible to build these models without using copyrighted material, such an argument has already started to be challenged.<sup>62</sup> Accordingly, whether LLMs could be developed by processing less or minimal publicly accessible online data is still an open question. Furthermore, it is an open secret that developers of these models are looking for more data sources rather than making their LLMs more efficient with less (personal) data processing. In fact, it has even started to be

speculated that they may run out of (high-quality) training data in the upcoming years, which already led these developers to look for alternatives to find data to train their models.<sup>63</sup> Given these considerations, it is very unlikely that OpenAI could pass the second step.

Lastly, regarding the third step, the CJEU underlined in the *Meta v Bundeskartellamt* case that the processing activities conducted by Meta had significant impacts on data subjects and, more importantly, highlighted that data subjects could not have reasonably expected their data to be processed by Meta.<sup>64</sup> When these considerations are applied to OpenAI's processing activities, it should be said that individuals whose personal data are scraped from publicly accessible online sources were unaware and could not have reasonably expected that their personal data would be used for training ChatGPT, especially considering the novelty of LLMs and the business models and products that are being developed on the mass processing of publicly accessible online data in general. Besides, there have been cases already reported where ChatGPT produced false information about individuals, which was sometimes defamatory.<sup>65</sup> This proves that OpenAI's processing activities conducted on 'information that is publicly available on the internet' could have significant adverse impacts on individuals. Furthermore, it should be noted that the CJEU highlighted in the *Meta v Bundeskartellamt* case that the scale of Meta's processing activities was so massive that it was liable to cause certain chilling effects, such as data subjects feeling as if their private lives were being continuously monitored.<sup>66</sup> Similar concerns exist also in the context at hand. The

60 Ibid paras 122–23.

61 ICO, Call for Evidence (n 50).

62 Kate Knibbs, 'Here's Proof You Can Train an AI Model Without Slurping Copyrighted Content' (WIRED, 20 March 2024) <<https://www.wired.com/story/proof-you-can-train-ai-without-slurping-copyrighted-content/>> accessed 31 May 2024. See also, 'There is a misconception that the principle of data minimization has no place in the context of artificial intelligence. However, data controllers have an obligation to limit the collection and otherwise processing of personal data to what is necessary for the purposes of the processing, avoiding indiscriminate processing of personal data. This obligation covers the entire lifecycle of the system, including testing, acceptance and release into production phases. Personal data should not be collected and processed indiscriminately.' European Data Protection Supervisor, 'Generative AI and the EUDPR. First EDPS Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems' (3 June 2024) 14.

63 Lakshmi Varanasi, 'Big Tech Needs to Get Creative as It Runs Out of Data to Train its AI Models. Here are Some of Its Wildest Solutions' (Business Insider, 7 April 2024) <<https://www.businessinsider.com/ai-training-data-source-solutions-openai-meta-google-2024-4?international=true&r=US&IR=T>> accessed 31 May 2024; Deepa Seetharaman, 'For Data-Guzzling AI Companies, the Internet Is Too Small' *The Wall Street Journal* (New York, 1 April 2024) <<https://www.wsj.com/tech/ai/ai-training-data-synthetic-openai-anthropic-9230f8d8>> accessed 31 May 2024. Aaron Mok, 'AI Giants Like OpenAI and Anthropic are Scrambling to Get Their Hands on Enough Data to Train Models'

(Business Insider, 1 April 2024) <<https://www.businessinsider.com/ai-giants-openai-anthropic-running-out-of-good-training-data-2024-4?international=true&r=US&IR=T>> accessed 31 May 2024.

64 C-252/21, Meta (n 48) para 123. See also, C-26/22 and C-64/22, UF, AB (n 48), para 80; EDPB, ChatGPT (n 29) para 16, page 6. See further, GDPR, recital 47:

(...) Such legitimate interest could exist for example where there is a relevant and appropriate relationship between the data subject and the controller in situations such as where the data subject is a client or in the service of the controller. At any rate the existence of a legitimate interest would need careful assessment including whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place. The interests and fundamental rights of the data subject could in particular override the interest of the data controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing.

65 Pranshu Verma and Will Oremus, 'ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused' *The Washington Post* (Washington DC, 5 April 2023) <<https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>> accessed 31 October 2023.

66 C-252/21, Meta (n 48) para 118.

race to mass processing of publicly accessible online data to train LLMs, and AI models in general, could lead to certain chilling effects on individuals, as they may tend to express themselves online less to prevent these models from inferring particularly detailed pictures of their private lives.<sup>67</sup> This is even more true for those in creative professions, as they may restrict the online disclosure of their work to prevent the developers from using them for training LLMs and gaining (financial) benefits without their permission. In fact, the lawsuits and complaints filed against these developers claiming several copyright and data protection infringements could only prove these concerns. Therefore, the fundamental rights and freedoms of data subjects, especially those in creative professions, could override the legitimate interest of OpenAI and/or other third parties when OpenAI processes the ‘information that is publicly available on the internet’ to train ChatGPT in the absence of effective (technical) measures to mitigate the risks posed on them.<sup>68</sup>

As it appears, OpenAI could pass the first step of the so-called three-step test of Article 6(1)(f) GDPR but may fail to do so regarding its second and third steps. Therefore, it is very unlikely that OpenAI could rely on Article 6(1)(f) GDPR for training ChatGPT with publicly accessible online data unless it alters its processing activities conducted on these data to adhere to the data protection principles, such as data minimization, and implement (technical) measures to mitigate the risks posed on the fundamental rights and freedoms of the data subjects concerned, especially those in creative professions. If the same line of reasoning is followed, the DPAs examining the lawfulness of these processing activities might also reach this conclusion as opposed to their Italian colleagues, who lifted the ban on ChatGPT after OpenAI switched the legal basis to Article 6(1)(f) GDPR for training purposes. However, even such a conclusion might be incorrect since the mass processing of

publicly accessible online data to train LLMs falls under the scope of Article 9 GDPR, as it involves sensitive data processing.

## Sensitive personal data processing by OpenAI when training ChatGPT with publicly accessible online data

In the EU data protection framework, certain types of personal data, such as those that reveal political opinions or concern the health or sexual orientation of data subjects, are considered sensitive, as their processing poses significant risks to data subjects’ fundamental rights and freedoms.<sup>69</sup> Accordingly, their processing is prohibited as per Article 9(1) GDPR,<sup>70</sup> unless it relies on one of the exceptions listed in Article 9(2) GDPR. While processing publicly accessible online data, OpenAI inevitably processes several types of sensitive data given the diverse nature of the information accessible on the Internet, such as data related to the political opinions or health status of individuals.<sup>71</sup> Therefore, at first glance, it should be inferred that the stricter protection regime provided by Article 9 GDPR instead of Article 6 GDPR should apply to the personal data processing conducted by OpenAI, at least for personal data relating to the categories listed in Article 9(1) GDPR.

However, this might also be an incorrect inference at this point. First, although the list provided in Article 9(1) GDPR is exhaustive, the CJEU has recently opened the door for a broad interpretation of what constitutes sensitive data. In the *OT*<sup>72</sup> judgment, the CJEU interpreted whether personal data that is liable to indirectly disclose the sexual orientation of a natural person constitutes sensitive data. By applying teleological interpretation, the CJEU provided a positive answer to this question and, therefore, included personal data, such as the name of the spouse, cohabitee, or partner of the data subject, that might (indirectly) reveal his or her

67 See further, Norwegian Consumer Council (n 11) 32.

68 See further, EDPB, ChatGPT (n 29) para 17, pages 6–7; ICO, Call for Evidence (n 50); EDPS, Generative AI (n 62) 7–12; Autoriteit Persoonsgegevens, ‘Scraping Door Particulieren en Private Organisaties’ (1 May 2024) 14.

69 GDPR, recital 51.

70 GDPR, art 9(1): ‘Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation shall be prohibited.’

71 EDPB, ChatGPT (n 29), para 15, page 6; Autoriteit Persoonsgegevens, ChatGPT (n 23); ‘TechSonar 2023-2024 Report’ (European Data Protection Supervisor, 4 December 2023), 8 <[https://www.edps.europa.eu/system/files/2023-12/23-12-04\\_techsonar\\_23-24\\_en.pdf](https://www.edps.europa.eu/system/files/2023-12/23-12-04_techsonar_23-24_en.pdf)> accessed 29 May 2024. In fact, there has already been a lawsuit filed against OpenAI in the United States claiming that OpenAI processed, among others,

medical data of the plaintiffs while training their models, see Grace Dean, ‘A lawsuit claims OpenAI stole ‘massive amounts of personal data,’ including medical records and information about children, to train ChatGPT’ (*Business Insider*, 29 July 2023) <<https://www.businessinsider.com/openai-chatgpt-generative-ai-stole-personal-data-lawsuit-children-medical-2023-6?international=true&r=US&IR=T>> accessed 1 November 2023. Furthermore, in a blog post published by Google regarding the privacy considerations of LLMs, it is mentioned that ‘Because these datasets can be large (hundreds of gigabytes) and pull from a range of sources, they can sometimes contain sensitive data, including personally identifiable information (PII) — names, phone numbers, addresses, etc, even if trained on public data.’ See, Nicholas Carlini, ‘Privacy Considerations in Large Language Models’ (Google Research, 15 December 2020) <<https://blog.research.google/2020/12/privacy-considerations-in-large.html>> accessed 1 November 2023.

72 *OT v Vyriausioji tarnybinės etikos komisija*, Case C-184/20, [2022] ECLI:EU:C:2022:601

sexual orientation within the scope of the protection regime of Article 9 GDPR.<sup>73</sup> While reaching this conclusion, it especially highlighted that when such information is published online, combining it with other information enables a particularly detailed picture of the private lives of data subjects to be drawn, increasing the seriousness of the interference with their fundamental rights and freedoms.<sup>74</sup>

This interpretation appears essential in determining to what extent personal data processed by OpenAI is subject to Article 9 GDPR regime. This is because personal data that is initially not considered sensitive could be combined with other (sensitive) personal data by ChatGPT to produce sensitive data about individuals as output while generating answers.<sup>75</sup> At this point, it should be mentioned that although OpenAI claims that it does not use data for building profiles of people,<sup>76</sup> it also expresses that ‘Our models may learn from personal information to (...) learn about famous people and public figures. This makes our models better at providing relevant responses.’<sup>77</sup> Such a statement could be understood as ChatGPT may use its training data to draw particularly detailed pictures of the private lives of individuals, at least for those belonging to certain groups, which may reveal sensitive information about them. Therefore, not only the personal data classified as sensitive by their nature but also those initially not considered sensitive could fall under the scope of Article 9 GDPR when training ChatGPT with publicly accessible online data. However, given the vast amount of personal data processed by OpenAI with this method, it would be practically impossible to discriminate between non-sensitive and (potentially) sensitive data at the time of their processing to determine which of those fall under the scope of Article 9 GDPR.

At this point, the *Meta v Bundeskartellamt* judgment should be revisited. In this case, the CJEU stated that when the processing in question contains both sensitive and non-sensitive data and such personal data is collected *en bloc* without the controller being able to

separate them at the time of the collection, these processing activities should be subjected to Article 9 GDPR.<sup>78</sup> This interpretation has serious implications for OpenAI. First, since OpenAI processes vast amounts of data scraped from the Internet to train its models, the training dataset of ChatGPT contains sensitive and non-sensitive personal data *en bloc*. Secondly, considering OpenAI states that it removes personal data from its training datasets ‘where feasible’, it appears that OpenAI cannot discriminate between (potentially) sensitive and non-sensitive personal data at the time of their collection. Therefore, *all* personal data that are publicly accessible online and processed by OpenAI to train ChatGPT falls under the scope of Article 9 GDPR.

In *Meta v Bundeskartellamt* judgment, the CJEU further stated that Article 9 GDPR applies irrespective of the fact that sensitive data in question relates to individuals other than the users of the controller, whether such information is correct, or the controller intended to obtain information that constitutes sensitive data.<sup>79</sup> Such conclusions, again, have serious implications for OpenAI. First, it is understood that the personal data of individuals other than ChatGPT users, whose personal data are scraped from publicly accessible sources online, are also subject to Article 9 GDPR. Secondly, data used for training ChatGPT that are inaccurate or even ChatGPT’s hallucinations, which correspond to the answers provided by ChatGPT that are inaccurate yet convincing,<sup>80</sup> could trigger the application of Article 9 GDPR regime. And thirdly, as long as sensitive data are processed, Article 9 GDPR applies regardless of OpenAI’s intention to teach its models ‘about the world, not private individuals’.<sup>81</sup> All these indicate that personal data processing done by OpenAI to train its models by using publicly accessible online data should be evaluated through Article 9 GDPR.<sup>82</sup>

This conclusion would lead processing activities by OpenAI to train ChatGPT with publicly accessible online data to be prohibited unless it relies on one of the exceptions provided in Article 9(2) GDPR. Among the

73 Ibid paras 117–28.

74 Ibid paras 99–101.

75 See also, ‘Multiple data sets can be combined in ways that cause harm: information that is not sensitive when spread across different databases can be extremely revealing when collected in a single place, and it can be used to make inferences about a person or population.’ Electronic Privacy Information Center, ‘Generating Harms: Generative AI’s Impact & Paths Forward’ (2023) 25.

76 ‘We use training information only to help our models learn about language and how to understand and respond to it. We do not and will not use any personal information in training information to build profiles about people, to contact them, to advertise to them, to try to sell them anything, or to sell the information itself.’ See, OpenAI, ChatGPT (n 25).

77 Ibid.

78 C-252/21, *Meta* (n 48) para 89.

79 Ibid paras 68–69.

80 ‘Today’s large language models predict the next series of words based on patterns they have previously seen, including the text input the user provides. In some cases, the next most likely words may not be factually accurate.’ OpenAI, AI Safety (n 13). See also, Court of Justice of the European Union, ‘Artificial Intelligence Strategy’ 4; Karen Weise and Cade Metz, ‘When A.I. Chatbots Hallucinate’ *The New York Times* (New York, 9 May 2023) <<https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html>> accessed 1 November 2023; Matt O’Brien, ‘Chatbots Sometimes Make Things Up. Is AI’s Hallucination Problem Fixable?’ (AP News, 1 August 2023) <<https://apnews.com/article/artificial-intelligence-hallucination-chatbots-chatgpt-falsehoods-ac4672c5b06e6f91050aa46ee731bcf4>> accessed 1 November 2023.

81 OpenAI, AI Safety (n 13)

82 See also, Autoriteit Persoonsgegevens, Scraping (n 68) 19.

exceptions provided therein, only two appear suitable for such purposes. The first is obtaining explicit consent from data subjects,<sup>83</sup> and the second is processing personal data that are manifestly made public by data subjects.<sup>84</sup> Since there is no hierarchy between these legal bases,<sup>85</sup> OpenAI could rely on either of them.

### Can OpenAI rely on Article 9(2)(a) GDPR?

According to Article 9(2)(a) GDPR, sensitive personal data could be processed based on the explicit consent of the data subject. While the GDPR defines consent in Article 4(11) GDPR, which requires consent to be freely given, specific, informed, and unambiguous, it falls short of defining what constitutes ‘explicit’ consent. Nevertheless, explicit consent is interpreted as the type of consent that fulfils the cumulative requirements enshrined in Article 4(11) GDPR and is accompanied by an express statement of the data subject.<sup>86</sup>

Freely given consent corresponds to data subjects’ genuine and free choice when providing consent and the ability to refuse or withdraw consent without detriment.<sup>87</sup> It also requires data subjects to be free from any inappropriate pressure or influence and considers whether there is a power imbalance between the data subject and the controller that might compromise the data subject’s free choice.<sup>88</sup> Specific consent corresponds to consent that is provided separately for different processing purposes.<sup>89</sup> Informed consent requires data subjects to receive information about the processing in an intelligible and easily accessible form with clear and plain language so they can make an informed decision.<sup>90</sup> Lastly, unambiguous consent requires the data subject’s clear and affirmative action.<sup>91</sup> Specifically,

Recital 32 GDPR states that while ticking a box when visiting a website or another statement or conduct that clearly indicates the data subject’s acceptance satisfies this condition, silence, pre-ticked boxes, or inactivity would fall short of doing so. Given these requirements, this condition is also considered closely related to the explicitness requirement,<sup>92</sup> which refers to the way the consent is expressed.<sup>93</sup> To that end, the EDPB mentions that the explicitness in online settings could be ensured by filling in an electronic form, sending an e-mail, uploading a scanned document with the signature, using an electronic signature, or ticking boxes on an explicit consent screen presented when visiting a website.<sup>94</sup>

As can be inferred from these explanations, explicit consent has strict requirements to be fulfilled. Therefore, it can be argued that relying on the explicit consent of data subjects as the legal basis for training ChatGPT by processing their publicly accessible online data does not appear feasible for OpenAI due to its practical and legal implications.<sup>95</sup> This is because, given that potentially millions of data subjects are subjected to these processing activities, it would be a very challenging, in fact impossible, task for OpenAI to obtain and demonstrate their explicit consent. For example, in order to comply with the informed consent condition, OpenAI needs to, among others, provide information about its processing activities to these individuals before processing their personal data to train ChatGPT. It would be practically impossible for OpenAI to do so given that it has no contact with these individuals to begin with, as their personal data are scraped from all different corners of the Internet.<sup>96</sup> Besides, even in such a

83 GDPR, art 9(2): ‘Paragraph 1 shall not apply if one of the following applies: (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject’.

84 GDPR, art 9(2): ‘Paragraph 1 shall not apply if one of the following applies: (e) processing relates to personal data which are manifestly made public by the data subject’.

85 European Data Protection Board, ‘Binding Decision 5/2022 on the dispute submitted by the Irish SA regarding WhatsApp Ireland Limited (Art. 65 GDPR)’ (5 December 2022), para 100, page 27.

86 See, European Data Protection Board, ‘Guidelines 03/2020 on the Processing of Data Concerning Health for the Purpose of Scientific Research in the Context of the COVID-19 Outbreak’ (30 April 2020), para 18, page 6; European Data Protection Board, ‘Guidelines 05/2020 on consent under Regulation 2016/679’ (13 May 2020), paras 92–93, pages 20–21.

87 GDPR, recital 42–43; See also, *Orange România SA v Autoritatea Națională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP)*, Case C-61/19, [2020] ECLI:EU:C:2020:901, para 41–50; EDPB, Consent (n 86) para 13, page 7.

88 GDPR, recital 43; EDPB, Consent (n 86) paras 13–14, pages 7–8.

89 GDPR, recital 43; *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband eV v Planet49*

*GmbH*, Case C-673/17, [2019] ECLI:EU:C:2019:801, paras 58–59; EDPB, Consent (n 86) paras 55–61, pages 13–15.

90 GDPR, art 7(2), art 7(3), and recital 42; C-61/19, *Orange România SA* (n 87) para 40; C-673/17, *Planet49 GmbH* (n 89) para 74.

91 ‘Only active behaviour on the part of the data subject with a view to giving his or her consent may fulfil that requirement.’ See, *Planet49 GmbH* (n 89) para 54.

92 Lee A Bygrave and Luca Tosoni, ‘Article 4(11). Consent’ in Christopher Kuner, Lee A Bygrave and Christopher Docksey (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press, Oxford, United Kingdom, 2020), 174, 185.

93 EDPB, Consent (n 86) para 93, pages 20–21.

94 *Ibid* paras 94–96, page 21.

95 In fact, the CJEU reached a similar conclusion regarding Google’s activities as a search engine operator, see, *GC, AF, BH, ED v Commission nationale de l’informatique et des libertés (CNIL)*, Case C-136/17, [2019] ECLI:EU:C:2019:773, para 62.

96 See also: ‘It does not appear possible to obtain valid consent in some cases. This is often the case when the controller collects publicly accessible data online or reuses an open dataset, especially given the lack of contact with the data subjects and the difficulty in identifying them. In these cases, where the conditions for obtaining valid consent are not met, the controller must rely on another, more appropriate, legal basis.’ CNIL (n 50).

case where OpenAI miraculously overcomes such practical hardships and manages to obtain and demonstrate the explicit consent of all these individuals, it would have to deal with further challenges as a result of relying on explicit consent as the legal basis for these processing activities. For instance, data subjects are entitled to withdraw their (explicit) consent at any time without that withdrawal retroactively affecting the lawfulness of the processing in question.<sup>97</sup> However, it is yet to be known whether and to what extent OpenAI could comply with such withdrawals to ensure that the personal data in question is no longer fed into the training datasets of ChatGPT.

Given such concerns about relying on Article 9(2)(a) GDPR as the legal basis for training ChatGPT with publicly accessible online data, OpenAI might seek the legal basis for these processing activities under another exception provided in Article 9(2) GDPR. At this point, it should also be remembered that although the Italian DPA asked OpenAI to switch its legal basis to consent or legitimate interest for training purposes, OpenAI decided to go for legitimate interest instead of consent. Perhaps such a decision was made to circumvent legal and practical hardships in obtaining the consent of these individuals. Accordingly, the same hesitance could be expected when obtaining their explicit consent is in question.

## Can OpenAI rely on Article 9(2)(e) GDPR?

As it appears that OpenAI cannot rely on Article 9(2)(a) GDPR for processing publicly accessible online data to train ChatGPT, the ‘manifestly made public by the data subject’ exception should be evaluated further to determine whether it could serve as the legal basis for such purposes. In fact, it would also be in line with the continuous claims of OpenAI that it processes information that is publicly accessible. At this point, it should also be recalled that the CJEU also hinted at such a direction in the *GC and Others v CNIL*<sup>98</sup> case by stating that this exception applies to search engine operators who also process publicly accessible online data *en masse*.<sup>99</sup> Nevertheless, the CJEU underlined that compliance with other obligations, specifically the principles set out

in Article 5 GDPR, should be upheld even in such cases.<sup>100</sup> Accordingly, evaluating whether this exception could provide a valid legal basis for OpenAI needs to be interpreted through this perspective by examining the relevant case law, official guidelines, and literature.

Enshrined in Article 9(2)(e) GDPR and Article 10(c) of the Law Enforcement Directive<sup>101</sup> in the EU data protection framework, the ‘manifestly made public by the data subject’ legal basis appears as one of the most neglected legal bases given the lack of legislative and judicial interpretation on it. Besides, very limited guidance has been provided by the EDPB and DPAs on this legal basis. And the very scarce literature dedicated to explaining what are the conditions of this legal basis mainly stated that the data subject’s intentional, voluntary, and positive act should be demonstrated to be able to rely on the ‘manifestly made public by the data subject’ exception when processing sensitive personal data.<sup>102</sup> Nevertheless, finally, in the *Meta v Bundeskartellamt* case, the CJEU clarified how this exception should be interpreted.

In this case, the CJEU first stated that this exception should be interpreted strictly, and personal data can be considered manifestly made public when data subjects have intended to make their personal data accessible to the general public explicitly and by clear affirmative action.<sup>103</sup> In this line, it was first acknowledged that a mere visit to a website or app where sensitive data are processed cannot be considered as manifestly making personal data public.<sup>104</sup> Then, the CJEU evaluated the situation in which data subjects enter information or, in other ways, interact with websites and apps that process sensitive data and underlined that whether that interaction is public should depend on the individual settings chosen by the data subject.<sup>105</sup> Therefore, if data subjects have an actual choice to do so and explicitly accepted to make their interactions, such as entering information on a website or app, accessible to the general public, instead of a more or less limited number of selected persons with full knowledge of the facts, their personal data could be considered manifestly made public by them.<sup>106</sup> On the other hand, if no such individual settings are available on the websites or apps with which the data subject interacts, such an evaluation should be made by

97 GDPR, art 7(3).

98 C-136/17, GC, AF, BH, ED (n 95).

99 Ibid para 63.

100 Ibid para 64; European Data Protection Board, 'Guidelines 8/2020 on the Targeting of Social Media Users' (7 July 2021), para 114, page 31.

101 Directive (EU) 2016/680 of the European Parliament and of the Council  
of 27 April 2016 on the protection of natural persons with regard to the  
processing of personal data by competent authorities for the purposes of  
the prevention, investigation, detection or prosecution of criminal offences  
or the execution of criminal penalties, and on the free movement of

such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L 119/89.

102 See, for instance, Edward S Dove and Jiahong Chen, 'What Does It Mean for a Data Subject to Make Their Personal Data 'Manifestly Public'? An Analysis of GDPR Article 9(2)(e)' (2021) 11 International Data Privacy Law 107, 122.

103 C-252/21, Meta (n 48) paras 76–77.

104 Ibid paras 78–79.

105 Ibid para 80.

106 Ibid paras 81–82.

inquiring whether data subjects have explicitly consented to their data being accessible by any person having access to that website or app, based on the express information provided to them.<sup>107</sup>

Currently, OpenAI mentions that: ‘Web pages crawled with the GPTBot user agent may potentially be used to improve future models and are filtered to remove sources that require paywall access, are known to primarily aggregate personally identifiable information (PII), or have text that violates our policies.’<sup>108</sup> Since no separation is mentioned between the websites that provide their users with such privacy settings and those that do not, whether OpenAI could rely on Article 9(2)(e) GDPR when training ChatGPT with publicly accessible online data should be evaluated regarding both cases.

### In cases where data subjects are presented with privacy settings to make their personal data publicly accessible

First, whether OpenAI could rely on Article 9(2)(e) GDPR for processing the personal data of data subjects who made their personal data publicly accessible through individual privacy settings should be evaluated. In such cases, the CJEU asks whether data subjects have clearly made their personal data publicly accessible with clear affirmative action and full knowledge of the facts.<sup>109</sup> Accordingly, three conditions that form this exception, namely, ‘manifestly made’, ‘public’, and ‘by the data subject’, should be evaluated through this perspective.

**‘Manifestly made’.** For personal data to be considered ‘manifestly made’ public, the CJEU requires an explicit, affirmative, and in advance choice of data subjects made with full knowledge of the facts.<sup>110</sup> An explicit and affirmative choice could be ensured with, for example, an opt-in mechanism embedded in the privacy settings, through which data subjects can choose to make their personal data accessible by (selected groups of) users of that website or the general public. This approach also aligns with the data protection by design and by default principle enshrined in Article 25 GDPR, as it specifically requires that ‘(...) by default personal data are not made accessible without the individual’s intervention to an indefinite number of natural persons.’<sup>111</sup> It is

observed that the EDPB also favoured this approach, as it listed the default settings of a platform among the conditions to be considered for relying on Article 9(2)(e) GDPR to process the personal data available on social media platforms.<sup>112</sup> Specifically, the EDPB mentioned that the default settings of a service that makes the personal data in question publicly accessible cannot lead this personal data to be classified as ‘manifestly made’ public.<sup>113</sup> Furthermore, the European Data Protection Supervisor (EDPS) also took into account whether a social media platform offers its users private and public options regarding the visibility of their accounts when deciding whether personal data processed from that platform is ‘manifestly made’ public.<sup>114</sup>

Based on these interpretations, it should be concluded that the personal data of users of the websites operating an opt-out policy, making personal data available on them automatically publicly accessible unless the data subject opposes this disclosure, cannot be considered as ‘manifestly made’ public. In other words, for this criterion to be met, OpenAI should demonstrate that data subjects whose personal data have been scraped from publicly accessible online sources have opted in to disclose their personal data through their privacy settings. However, the proportion of publicly accessible online data that complies with this requirement is quite debatable, given the vast number of websites that do not comply with this obligation. In fact, it should be noted that even Meta has been slapped on the wrist not once but twice for infringing this obligation by the Irish DPA in cases related to Instagram and Facebook, while the latter case even had a scraping element regarding the publicly exposed personal data.<sup>115</sup> Thus, it becomes very questionable to what extent the publicly accessible online data that OpenAI processes to train ChatGPT comply with this criterion.

Furthermore, even in cases where data subjects choose to make their personal data publicly accessible, the CJEU requires such a decision to be made with full knowledge of the facts. Although the *Meta v Bundeskartellamt* judgment fails to explain further what should be understood from the ‘facts’, by applying teleological interpretation, it can be argued that it corresponds to the potential consequences of disclosing the

107 Ibid para 83.

108 OpenAI, GPTBot (n 14)

109 C-252/21, *Meta* (n 48) paras 77–82.

110 C-252/21, *Meta* (n 48) para 85.

111 GDPR, art 25(2).

112 EDPB, Social Media (n 100) para 127, page 35.

113 European Data Protection Board, ‘Guidelines 05/2022 On the Use of Facial Recognition Technology in the Area of Law Enforcement’, V 2.0, (26 April 2023) para 76, page 22.

114 European Data Protection Supervisor, ‘EDPS Supervisory Opinion on the Use of Social Media Monitoring for Epidemic Intelligence Purposes by The European Centre for Disease Prevention and Control (‘ECDC’)’ (9 November 2023) para 64, pages 17–18.

115 Case In-20-7-4, *Meta Platforms Ireland Limited and Instagram*, Data Protection Commission [2022] <[https://www.edpb.europa.eu/system/files/2022-09/in-20-7-4\\_final\\_decision\\_-\\_redacted.pdf](https://www.edpb.europa.eu/system/files/2022-09/in-20-7-4_final_decision_-_redacted.pdf)>; Case IN-21-4-2, *Meta Platforms Ireland Ltd*, Data Protection Commission [2022], para 182, pages 59–60 <[https://www.dataprotection.ie/sites/default/files/uploads/2022-12/Final%20Decision\\_IN-21-4-2\\_Redacted.pdf](https://www.dataprotection.ie/sites/default/files/uploads/2022-12/Final%20Decision_IN-21-4-2_Redacted.pdf)>

personal data in question to the public.<sup>116</sup> In practice, data subjects learn about these ‘facts’ and decide whether to opt-in to make their personal data accessible to the public by reading documents provided to them by the websites they are interacting with, such as terms of services and privacy policies. However, these documents have long been criticized for, among others, their length, unintelligible language, and take-it-or-leave-it approach.<sup>117</sup> Furthermore, they contain vague statements due to their heavy reliance on modal verbs, for example can, could, may, and might, creating uncertainty regarding the purposes and possible further uses of personal data.<sup>118</sup> Likewise, these documents may nudge data subjects towards certain options even though data subjects should maintain the highest degree of autonomy possible regarding their choices.<sup>119</sup> Stemming from these concerns, the EDPB mentioned that the visibility of the information where the data subjects are informed about the ‘facts’ constitutes a condition in determining whether the personal data in question has been ‘manifestly made’ public.<sup>120</sup> Accordingly, if the ‘facts’ are, for example, hidden in a lengthy privacy policy that nudges data subjects to disclose their personal data to the public, there can be no ‘manifestly made’ choice of the data subject.

Even where data subjects have been properly informed about the ‘facts’ and autonomously opt to disclose their personal data to the public, certain practices may still compromise their ‘manifestly made’ choices. For example, it is well-known that the documents informing data subjects about the ‘facts’ are subject to frequent changes that are not adequately communicated to data subjects. These silent changes, however, may compromise the data subjects’ ‘in advance’ choices. This is because, due to such changes, for instance, the categories of personal data disclosed to the public may be

extended beyond those initially agreed by the data subjects. In fact, the US Federal Trade Commission very recently warned companies, especially AI companies, not to quietly change their terms of services to allow the sharing of their consumers’ data with third parties or using that data for training AI models.<sup>121</sup> Besides, even when data subjects have the full knowledge of the ‘facts’, they may still be willing to disclose their personal data for reasons such as the network effect or to be able to use a service for free, which would, again, compromise their ‘manifestly made’ choices as their autonomy would be infringed in such cases.<sup>122</sup>

Overall, certain practices may prevent data subjects from having full knowledge of the facts or compromise their ‘manifestly made’ decisions to disclose their personal data to the public. Therefore, whether and how OpenAI filters and processes only the personal data of data subjects who have ‘manifestly made’ their personal data public requires further explanation. In fact, even in such cases, OpenAI may not be able to rely on Article 9 (2)(e) GDPR to process these data to train ChatGPT. This is because, as rightly put by the EDPB, ‘The word ‘manifestly’ implies that there must be a high threshold for relying on this exemption.’<sup>123</sup> To that end, I argue that the reasonable expectations of data subjects should also be taken into account when determining whether they ‘manifestly made’ their personal data public.

Indeed, the reasonable expectation of the data subject is an important point of reference regarding the assessment of the lawfulness of a processing activity, proven by the fact that the European legislator explicitly referred to it in Recitals 47 and 50 GDPR in this context. Accordingly, it is observed that, for example, the DPAs scrutinizing Clearview AI specifically emphasized that data subjects could not have reasonably expected their images to be processed to create a facial recognition app

116 See, GDPR, recital 39; C-61/19, *Orange România SA* (n 87) para 40.

117 See, for instance, Isabel Wagner, ‘Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021’ (2023) 26 *ACM Transactions on Privacy and Security* 1; Aleecia M McDonald and Lorrie F Cranor, ‘The Cost of Reading Privacy Policies’ (2008) 4 *Journal of Law and Policy for the Information Society* 543; Midas Nouwens and others, ‘Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating Their Influence’ in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 1; Kevin Litman-Navarro, ‘We Read 150 Privacy Policies. They Were an Incomprehensible Disaster’ *The New York Times* (New York, 12 June 2019) <<https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html>> accessed 1 November 2023.

118 See further, Article 29 Working Party, ‘Guidelines on Transparency under Regulation 2016/679’ (11 April 2018) WP260, para 13, page 9.

119 European Data Protection Board, ‘Guidelines 4/2019 on Article 25 Data Protection by Design and by Default’ (20 October 2020) para 70, pages 18–19.

120 EDPB, *Social Media* (n 100) para 127, page 35.

121 ‘AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive’ (*Federal Trade Commission*, 13 February

2024), <<https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive>> last accessed on 30 May 2024. See further, Kali Hays, ‘A Long List of Tech Companies are Rushing to Give Themselves the Right to Use people’s Data to Train AI’ (*Business Insider*, 13 September 2023) <<https://www.businessinsider.com/tech-updated-terms-to-use-customer-data-to-train-ai-2023-9?international=true&r=US&IR=T>> accessed 30 May 2024; Cade Metz and others, ‘How Tech Giants Cut Corners to Harvest Data for A.I.’ *The New York Times* (New York, 6 April 2024) <[https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html?mc\\_cid=14d2fad1ae&mc\\_eid=dd0e87f701](https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html?mc_cid=14d2fad1ae&mc_eid=dd0e87f701)> accessed 31 May 2024.

122 Mateja Durovic and Franciszek Lech, ‘A Consumer Law Perspective on the Commercialization of Data’ (2021) 29 *European Review of Private Law* 701, 714; Bart Custers and others, ‘The Role of Consent in an Algorithmic Society—Its Evolution, Scope, Failings and Re-conceptualization’ in Eleni Kosta, Ronald Leenes and Irene Kamara (eds), *Research Handbook on EU Data Protection* (Edward Elgar Publishing, Cheltenham, United Kingdom, 2022) 455, 467. See also, EDPB, *Binding Decision* (n 85) para 119, page 31.

123 EDPB, *Social Media* (n 100) para 127, page 35.

while determining that Clearview AI's processing activities were unlawful.<sup>124</sup> Furthermore, even though the EDPS stated that the European Center for Disease Prevention and Control (ECDC) could rely on the 'manifestly made public by the data subject' legal basis to process the personal data of Twitter users with a public account for epidemic intelligence purposes, it nevertheless noted that: '(...) it is debateable whether such a Twitter user would reasonably expect that its public post containing special category data would subsequently be further processed by the ECDC via epitweetr for purposes of detecting and preventing potential outbreaks of communicable diseases.'<sup>125</sup> Moreover, very recently, Advocate General (AG) Rantos also, indirectly, underlined the reasonable expectations of data subjects regarding processing their sensitive personal data based on Article 9(2)(e) GDPR. In his opinion on the *Maximilian Schrems v Meta Platforms Ireland Limited*<sup>126</sup> case, where, among others, whether a statement made by an individual regarding his sexual orientation in a panel discussion could enable the processing of such data for personalized advertising purposes is discussed, the AG first stated that by making statements about his sexual orientation in a panel discussion that attracted public interest, was open for the press, and broadcasted live, the data subject manifestly made his data relating to his sexual orientation public as per Article 9(2)(e) GDPR.<sup>127</sup> However, the AG opined that such a disclosure does not, in itself, enable the processing of these data for personalized advertising purposes by underlining the purpose limitation principle, which is enshrined in Article 5(1)(b) GDPR and requires data controllers to process personal data for specified, explicit, and legitimate purposes.<sup>128</sup>

Similarly, it can be argued that when disclosing their personal data to the public, data subjects could not have reasonably expected their data to be used for training LLMs. This conclusion stems not only from the novelty of the techniques developed to utilize the publicly accessible online data to train these models but also from their invisibility. For example, it has been reported that Google, rather quietly and discreetly, updated its privacy policy to mention that

publicly accessible data scraped from the Internet may be used for, among others, training its LLMs.<sup>129</sup> In fact, it has been observed that more and more companies are updating their privacy policies to allow themselves to use their customers' and/or publicly accessible online data to train AI models, including LLMs.<sup>130</sup> These invisible processing activities, however, are liable to lead data subjects to lose control over their personal data and their (subsequent) uses, which may infringe on their fundamental rights and freedoms.<sup>131</sup> These concerns are even more relevant for data subjects who are in creative professions since their reasonable expectation was merely to increase the visibility of their work by disclosing them to the public and not for them to be used for training LLMs, as evidenced by the lawsuits and complaints the developers of these models face globally. Accordingly, OpenAI and other developers should also consider the reasonable expectations of data subjects when determining whether they have 'manifestly made' their personal data public. However, as it appears, they rather disregard these expectations, which could prevent them from relying on Article 9(2)(e) GDPR to process publicly accessible online data to train their models.

**'Public'.** OpenAI could rely on Article 9(2)(e) GDPR to process publicly accessible online data as long as they are classified as 'public'. In other words, the personal data in question should have been disclosed to the 'public' because of the data subject's manifestly made decision. However, the GDPR falls short of defining what constitutes 'public'. Only in Article 25(2) GDPR, it is mentioned that '(...) by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.' This signifies that the term 'indefinite number of natural persons' can guide understanding of what 'public' means in online settings in the EU data protection framework. The CJEU jurisprudence also appears to be supporting this claim. This is because when the CJEU refers to the 'public' in online settings, it uses 'unrestricted number of people', 'indefinite number of people', 'unlimited number of persons', and 'all Internet users'

124 Pathak (n 4) 421–22.

125 EDPS, ECDC (n 114) para 64, page 18.

126 *Maximilian Schrems v Meta Platforms Ireland Limited*, Case C-446/21, ECLI:EU:C:2024:366.

127 Opinion of Advocate General Rantos in *Maximilian Schrems v Meta Platforms Ireland Limited*, Case C-446/21, ECLI:EU:C:2024:366, paras 40–44.

128 Ibid paras 45–46.

129 Jess Weatherbed, 'Google Confirms It's Training Bard on Scraped Web Data, Too' (*The Verge*, 5 July 2023) <<https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping>> last

accessed on 31 October 2023. See further, Sara Morrison, 'The Tricky Truth about How Generative AI Uses Your Data' (*Vox*, 27 July 2023) <<https://www.vox.com/technology/2023/7/27/23808499/ai-openai-google-meta-data-privacy-nope>> accessed 30 May 2024; Sarah Perez, 'X's Privacy Policy Confirms It Will Use Public Data to Train AI Models' <<https://techcrunch.com/2023/09/01/xs-privacy-policy-confirms-it-will-use-public-data-to-train-ai-models/>> accessed 30 May 2024.

130 Hays (n 121)

131 ICO, Call for Evidence (n 50). See also, C-184/20, OT (n 72) paras 103–05.

interchangeably.<sup>132</sup> Moreover, specifically in the contexts of data and copyright protection, it checks whether any restrictive measure has been applied to prevent everyone from accessing personal data or the work in question to determine its ‘public’ status. In fact, if no restrictive measures are applied to prevent everyone from accessing the personal data in question, the CJEU assigns ‘public’ status to it even when its dissemination is directed to a specific institution instead of the general public.<sup>133</sup> Consequently, it can be inferred that if no restrictive measures are applied and, therefore, the personal data in question is accessible by any Internet user instead of a more or less restricted number or category of users of the website it is posted on, such personal data should be classified as ‘public’.

The EDPB and the EDPS have also followed this line of reasoning, as both underlined that personal data should be accessible by everyone for it to be processed based on the ‘manifestly made public by the data subject’ legal basis.<sup>134</sup> Specifically, while the EDPB listed the accessibility of a webpage as one of the elements to consider when determining the ‘public’ status of the personal data in question, the EDPS assigned ‘public’ status to Tweets posted from public accounts.<sup>135</sup> Several DPAs have also issued decisions by following this approach.<sup>136</sup>

This interpretation would allow OpenAI to process the personal data that are accessible by any Internet user based on Article 9(2)(e) GDPR, provided that other conditions of this provision are also met. However, such a clear-cut solution could lead to unfair outcomes as it fails to understand how private and public spheres are defined in online discourses today. For example, does this interpretation mean an Instagram post published on a closed account

with 20 million followers is considered ‘private’, but a post published without any hashtags on an open account with only five followers is considered ‘public’? Hence, a contextual approach again appears necessary to determine the ‘public’ status of personal data.

In fact, this contextual approach has also been supported by the EDPB, EDPS, DPAs, and the literature. For example, on the processing of personal data of social media users, the EDPB underscored that the nature of the social media platform should be taken into account when relying on Article 9(2)(e) GDPR.<sup>137</sup> Specifically, it should be questioned whether the platform aims to connect the data subjects with their close friends and family and create intimate relations or allow them to interact with people beyond their (immediate) social cycles.<sup>138</sup> The ICO also used social media to describe this contextuality. It stated that it would be hard to consider personal data manifestly made ‘public’ if the post is directed to family and friends, even if it is accessible by the public due to the default privacy settings of that platform.<sup>139</sup> The EDPS illustrated this contextuality by stating, ‘Publishing personal data in a biography or an article in the press is not the same as posting a message on a social media page’.<sup>140</sup> Furthermore, the ICO mentioned that personal data should be realistically accessible by the public and underlined that ‘information is not necessarily public just because you have access to it’.<sup>141</sup> Following a similar line of reasoning, the Norwegian DPA found that Grindr<sup>142</sup> users did not manifestly make their personal data ‘public’ by using the app since their profiles were shown only to a limited number of other users based on their location.<sup>143</sup> Lastly, Dove and Chen mentioned that if access to data

132 *Bodil Lindqvist*, Case C-101/01, [2003] ECR I-12971 (ECLI:EU:C:2003:596), para 58; C-73/07, *Tietosuojavaltuutettu v Satakunnan Markkinapörssi Oy and Satamedia Oy* (n 38) para 44; *Nils Svensson, Sten Sjögren, Madelaine Sahlman and Pia Gadd v Retriever Sverige AB*, Case C-466/12, [2014] ECLI:EU:C:2014:76, para 26; *GS Media BV v Sanoma Media Netherlands BV, Playboy Enterprises International Inc. and Britt Geertruida Dekker*, Case C-160/15, [2016] ECLI:EU:C:2016:644, paras 42–48; *Stichting Brein v Jack Frederik Willems*, Case C-527/15, [2017] ECLI:EU:C:2017:300, para 48; *Land Nordrhein-Westfalen v Dirk Renckhoff*, Case C-161/17, [2018] ECLI:EU:C:2018:634, para 37; *BY v CX*, Case C-637/19, [2020] ECLI:EU:C:2020:863, para 26; *VG Bild-Kunst v Stiftung Preußischer Kulturbesitz*, Case C-392/19, [2021] ECLI:EU:C:2021:181, para 37; *Frank Peterson v Google LLC, YouTube Inc., YouTube LLC, Google Germany GmbH and Elsevier Inc. v Cyando AG*, Cases C-682/18 and C-683/18, [2021] ECLI:EU:C:2021:503, paras 72–75; *WM, Sovim SA v Luxembourg Business Registers*, Cases C-37/20 and C-601/20, [2022] ECLI:EU:C:2022:912, para 42; C-184/20, *OT* (n 72) para 102; C-252/21, *C-252/21, Meta* (n 48) para 85.

133 See, *Monika Esch-Leonhardt, Tillmann Frommhold and Emmanuel Larue v European Central Bank*, Case C-320/02, [2004] ECLI:EU:T:2004:45. See further, *Tietosuojavaltuutettu*, Case C-25/17, [2018] ECLI:EU:C:2018:551.

134 European Data Protection Supervisor, ‘A Preliminary Opinion on Data Protection and Scientific Research’ (6 January 2020) 19; Article 29 Data Protection Working Party, ‘Opinion on Some Key Issues of the Law Enforcement Directive (EU 2016/680)’ (29 November 2017) 10.

135 EDPB, *Social Media* (n 100) para 127, page 35; EDPS, *ECDC* (n 114) para 63, page 17.

136 See, for instance, Case 2020010552, *Persónuvernd* [2021] <[https://gdprhub.eu/index.php?title=Pers%C3%B3nuvernd\\_\(Iceland\)\\_-\\_no.\\_2020010552](https://gdprhub.eu/index.php?title=Pers%C3%B3nuvernd_(Iceland)_-_no._2020010552)> accessed 1 November 2023. See also, *Case C/05/368427/KG ZA 20-106, X v Y* [2020] ECLI:NL:RBGEL:2020:2521 <[https://gdprhub.eu/index.php?title=Rb.\\_Gelderland\\_-\\_C/05/368427](https://gdprhub.eu/index.php?title=Rb._Gelderland_-_C/05/368427)> accessed 1 November 2023.

137 EDPB, *Social Media* (n 100) para 127, page 35.

138 *Ibid.*

139 Information Commissioner’s Office, ‘What are the Conditions for Processing?’ (26 October 2023) <<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-are-the-conditions-for-processing/>> accessed 1 November 2023.

140 EDPS, *Scientific research* (n 134) 19.

141 ICO, *Conditions* (n 139); ‘However, the mere fact that personal data is publicly accessible does not imply that ‘the data subject has manifestly made such data public’ see EDPB, *ChatGPT* (n 29) para 18, page 7.

142 A social networking app for gay, bi, trans, and queer people that has millions of daily users that are being shown to each other based on their location, see, ‘About’ (*Grindr*) <<https://www.grindr.com/about>> accessed 30 May 2024.

143 ‘Grindr is a LGBTQ+ dating and social networking app, used for creating intimate relations or connecting with other users in the LGBTQ+’

requires a disproportionate and/or resource-intensive effort, it cannot be considered ‘public’.<sup>144</sup>

If these contextual considerations are taken into account, the pool of personal data that is in the ‘public’ domain and, therefore, could be processed by OpenAI to train ChatGPT, provided that other conditions of Article 9(2)(e) GDPR are also met, would be significantly limited. This is because it would oblige OpenAI to take into account, for example, whether the personal data in question are directed to close family and friends of the data subject and/or is realistically accessible while determining its ‘public’ status. However, the current understanding of the CJEU on what constitutes ‘public’ in online settings appears to allow OpenAI to process personal data that is accessible by any Internet user to train ChatGPT based on Article 9(2)(e) GDPR, provided that other conditions of this provision are also met. Nevertheless, it should be remembered that, as underlined by the CJEU and the EDPS, even in such cases, compliance with other obligations outlined in the GDPR, especially respecting data protection principles enshrined in Article 5 GDPR, should be ensured.<sup>145</sup>

**‘By the data subject’.** This condition appears to be the most straightforward one, requiring data subjects themselves to disclose their personal data to the public. Accordingly, if a data subject did not opt-in to disclose his/her personal data to the public but it was disclosed by third parties, that personal data could not be considered manifestly made public ‘by the data subject’.<sup>146</sup> Such cases can occur due to, for example, data leaks, technical glitches, or even changes in the disclosure policies of websites that expand the categories of personal data disclosed to the public without informing the users about such a change. If the personal data in question are disclosed to the public under such circumstances, it

cannot be processed based on Article 9(2)(e) GDPR, as its disclosure is ensured by others instead of the data subject.

This conclusion has serious implications for OpenAI. This is because it uses various online sources to train ChatGPT and, as underlined by the ICO, ‘the internet also contains information that was not placed there by the person to whom it relates’.<sup>147</sup> In fact, it has already been reported that AI training datasets created by scraping publicly accessible online (personal) data contain leaked (personal) data.<sup>148</sup> Furthermore, it should be recalled that OpenAI states that ‘models may learn from personal information to (...) learn about famous people and public figures’.<sup>149</sup> This statement implies that sources in which information about data subjects are disclosed by third parties, such as news articles, are also used by OpenAI to train ChatGPT. However, verifying whether data subjects themselves are responsible for the disclosure of the information presented therein is often impossible in such cases, while, as stated by the ICO, data controllers should be ‘confident that it was the individual themselves who actively chose to make their special category data public’.<sup>150</sup> Furthermore, the situation becomes even more problematic in cases where a data subject, for example, deletes a Tweet or switches the account status from public to private, but the information presented in that Tweet remains available on other sources, which are then used by OpenAI to train ChatGPT.<sup>151</sup> Similarly, the information presented on sources such as publicly accessible social media posts or personal blogs may contain personal data of individuals other than the owner of those accounts. In such cases, while the personal data of the owners of those accounts could be considered as manifestly made public ‘by the data subject’, provided that other conditions of Article 9 (2)(e) GDPR are also met, the same cannot be said for the personal data of other individuals unless it is verified that they were responsible for the disclosure of

community. In our view, there is a distinct difference between making information solely available to a community of peers, and making information available to the general public. (...) even though Grindr has approximately x million daily users, the data subject's Grindr profile would in practice only be shown to a limited amount of users, and most of these would be Grindr users near the user's actual or chosen location', see, Case 20/02136-18, *Datatilsynet* [2021], pages 46–47 <<https://www.datatilsynet.no/contentassets/8ad827efefcb489ab1c7ba129609edb5/administrative-fine—grindr-llc.pdf>> accessed 1 November 2023.

144 Dove and Chen (n 102) 122.

145 C-136/17, GC, AF, BH, ED (n 95) para 64; EDPS, ECDC (n 114) para 65, page 18.

146 C-252/21, Meta (n 48) para 75; EDPB, Social Media (n 100) para 127, page 35; EDPB, Facial Recognition (n 113), para 75, page 21; EDPS, ECDC (n 114) para 63, page 17.

147 ICO, Call for evidence (n 50)

148 Benj Edwards, 'Artist Finds Private Medical Record Photos in Popular AI Training Data Set' (*Ars Technica*, 21 September 2022) <<https://arstechnica.com/ai/2022/09/ai-artists-private-medical-photos/>>

[ica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/](https://www.ica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/) accessed 30 May 2024.

149 OpenAI, ChatGPT (n 25)

150 ICO, Conditions (n 139)

151 See, 'In addition, the use of the present tense in the phrase 'are manifestly made public' appears to purposively exclude situations where a data subject published data (eg through an injudicious public social networking post) but has then taken active steps to reassert their privacy (eg by, for example, deleting this post), at least where a sufficient time has elapsed such that the initial publication can reasonably be considered to be in the past.' David Erdos, 'Special, Personal and Broad Expression: Exploring Freedom of Expression Norms under the General Data Protection Regulation' (2021) 40 Yearbook of European Law, 398, 423. See further, Karen Hao, 'Deleting Unethical Data Sets Isn't Good Enough' (*MIT Technology Review*, 13 August 2021) <<https://www.technologyreview.com/2021/08/13/1031836/ai-ethics-responsible-data-stewardship/>> accessed 1 November 2023.

that information. Although OpenAI states that ‘Web pages crawled with the GPTBot user agent ( ... ) are filtered to remove sources that require paywall access, are known to primarily aggregate personally identifiable information (PII), or have text that violates our policies’,<sup>152</sup> currently, there seems to be no reference to these specific concerns. Therefore, until OpenAI’s algorithms are smart enough to detect which personal data have been disclosed by data subjects themselves, it will be questionable whether the ‘by the data subject’ condition is met when OpenAI relies on Article 9(2)(e) GDPR to train ChatGPT with the publicly accessible online data.<sup>153</sup>

Given all these explanations, it appears that OpenAI could only satisfy the ‘public’ condition of Article 9(2) (e) GDPR when it processes, for training ChatGPT, the personal data of individuals who are presented with privacy settings to disclose their personal data to the public. On the other hand, it is very unlikely these data will meet the high threshold of the ‘manifestly made’ condition, and, in certain cases, they are not disclosed ‘by the data subject’. Hence, the pool of personal data that is publicly accessible online and could be processed by OpenAI to train ChatGPT by relying on Article 9(2)(e) GDPR appears limited. This conclusion, however, may lead to questioning the lawfulness of OpenAI’s processing activities to train ChatGPT by using publicly accessible online data, as, to the best of our knowledge, there has been no clarification made by OpenAI that only this limited amount of personal data has been used for such purposes, a concern that applies for other LLM developers, too.

In the absence of such privacy settings

According to the CJEU, if data subjects are not presented with individual privacy settings to disclose their personal data to the public, Article 9(2)(e) GDPR could be used as the legal basis to process their data if they have explicitly consented to disclose their data to any person having access to the website or app in question based on the express information they have received.<sup>154</sup> This requirement also stems from the data protection by design and by default principle enshrined in Article 25 GDPR, which requires personal data not to be disclosed to an indefinite number of persons without the data subject's intervention.<sup>155</sup> Accordingly, where data subjects are not presented with privacy settings

regarding the disclosure of their sensitive personal data, such an intervention manifests itself as the explicit consent of these data subjects. Therefore, OpenAI should demonstrate that these individuals have explicitly consented to disclose their data to the public based on the express information they received to be able to process their data to train ChatGPT by relying on Article 9(2) (e) GDPR.

However, it is not an easy task since, in such cases, whether the explicit consent requirements were met is highly questionable, let alone the practical impossibility of OpenAI verifying this compliance. This is because, like the concerns defined in earlier sections, several practices may compromise the data subject's explicit consent. For example, freely given consent requires data subjects to enjoy a high degree of autonomy when giving their consent; therefore, any compromising acts, such as nudging, should be avoided. In practice, however, this condition is not met in many cases since, for instance, data controllers use dark patterns to obtain the consent of their users.<sup>156</sup> Besides, data subjects often provide their consent by clicking on general options, such as 'Accept all'. These practices force data subjects to accept all the processing purposes in bulk, which contradicts the specific consent requirement.<sup>157</sup> The informed consent requirement could also be compromised by the abovementioned shortcomings of privacy policies through which data subjects are informed about the processing activities and their implications. Yet, it should be noted that complying with this condition is especially crucial since the CJEU requires data subjects to provide their explicit consent based on the express information they receive. Lastly, in connection with the explicitness requirement, the unambiguous consent requirement could also be compromised when websites collect the consent of their users through pre-ticked boxes or consider inactivity or silence as consent.

Given these concerns, it can be argued that websites that do not provide their users with privacy settings regarding the disclosure of their personal data to the public but obtain their explicit consent for this disclosure by fully complying with its strict requirements could only represent a small percentage in practice. Whether and how OpenAI can identify these websites and their users who have explicitly consented to the public disclosure of their personal data is open for debate. Besides, even for these websites, as explained in earlier sections,

152 OpenAI, GPTBot (n 14)

153 See further, Bert-Jaap Koops, 'Police Investigations in Internet Open Sources: Procedural-law Issues' (2013) 29 Computer Law & Security Review 654, 665.

154 C-252/21, Meta (n 48) para 83.

155 GDPR art 25(2).

156 See eg, 'Provvedimento prescrittivo e sanzionatorio nei confronti di Ediscom S.p.A. - 23 febbraio 2023 [9870014]' (*Garante per la protezione dei dati personali*, 17 April 2023) <<https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870014>> accessed 1 November 2023.

157 See eg, Datatilsynet (n 143) 17–20.

it can be argued that the reasonable expectations of their users when disclosing their personal data to the public might prevent OpenAI from processing their personal data based on Article 9(2)(e) GDPR to train ChatGPT. Moreover, OpenAI should still verify that it was the data subjects themselves who ensured the public disclosure of the processed personal data due to the 'by the data subject' condition. These concerns, however, would seriously limit, again, the pool of personal data that OpenAI can process to train ChatGPT by relying on Article 9(2)(e) GDPR. As mentioned, OpenAI states that 'Web pages crawled with the GPTBot user agent (...) are filtered to remove sources that require paywall access, are known to primarily aggregate personally identifiable information (PII), or have text that violates our policies'.<sup>158</sup> Yet, to the best of our knowledge, there has been no specific reference made by OpenAI that it filters and removes the personal data collected from websites that do not comply with these requirements when training ChatGPT with publicly accessible online data. Therefore, further clarification and scrutiny are required to ensure that only this limited amount of personal data that complies with the requirements of Article 9(2)(e) GDPR is being used for training ChatGPT, a concern that also applies to other developers.

In short, regardless of whether data subjects are presented with privacy settings to publicly disclose their personal data or not, it appears that only a small proportion of personal data that are publicly accessible online comply with the requirements of Article 9(2)(e) GDPR for them to be processed for training LLMs by relying on this legal basis. However, developers of these models appear to continue their race to process publicly accessible online data *en masse* and indiscriminately for training their LLMs. Accordingly, these practices could be called 'data trawling' given their similarities with the 'bottom trawling' technique used in fisheries.

'Bottom trawling' is a technique used for catching large numbers of (different types of) fish at once by dragging big, heavy nets on the sea floor. It is severely criticized for two main reasons. First, these nets catch large amounts of fish and other marine species indiscriminately without separating, or being able to separate, the targeted population from the others, which can also include protected populations. Secondly, the nets

used for this technique can damage the seafloor, disturbing the habitat of many species living in those ecosystems. Therefore, this technique threatens ocean life and biodiversity, which has led to its ban or restriction in several marine areas worldwide.<sup>159</sup>

'Data trawling' for training LLMs, and AI models in general, mirrors this technique and its implications. First, even though 'data trawling' may occur in the 'public' domain, corresponding to the cases where 'bottom trawling' takes place in the territories allowing this technique, it allows developers to indiscriminately process several types of (personal) data in bulk. The developers can only filter the personal data that comply with the requirements of Article 9(2)(e) GDPR only after the 'data trawling' has already been conducted and 'where feasible'.<sup>160</sup> Secondly, 'data trawling' relies on scraping techniques such as crawlers, which are generally used by certain actors such as search engines, researchers, and academics. Therefore, it fundamentally challenges some of the Internet's (the online habitat) unwritten rules.<sup>161</sup> Hence, as explained in earlier sections, 'data trawling' for training LLMs is liable to cause certain chilling effects in the online world and infringe on the fundamental rights and freedoms of individuals whose personal data are subjected to this practice, especially those in creative professions. Consequently, these concerns raise questions about the legality and legitimacy of 'data trawling' for training LLMs, which also applies to developers using the same processing technique to train other AI models.

## A possible way forward?

As demonstrated, OpenAI may not easily rely on the usual suspects, namely Article 9(2)(a) and (e) GDPR, for processing publicly accessible online data to train ChatGPT. Hence, OpenAI may try to circumvent the prohibition enshrined in Article 9(1) GDPR in another way that will allow it to continue these processing activities. While the GDPR has provisions allowing certain types of processing activities to rely on several exemptions and derogations, such as Article 2(2)(c) GDPR, which allows processing activities carried out by a natural person in the course of purely personal or household activity to escape from the GDPR's material scope, and Article 85(2) GDPR, which is related to processing

158 OpenAI, GPTBot (n 14)

159 Ellie Hooper, 'What is Bottom Trawling and Why is it Bad for the Environment?' (*Greenpeace*, 11 April 2020) <<https://www.greenpeace.org/aotearoa/story/what-is-bottom-trawling-and-why-is-it-bad-for-the-environment/>> accessed 31 May 2024; Jamie Hailstone, 'Why Calls to Ban Bottom Trawling are Growing' *Forbes* (New Jersey, 9 August 2023) <<https://www.forbes.com/sites/jamiehailstone/2023/08/09/why-calls-to-ban-bottom-trawling-are-growing/>> accessed 31 May 2024.

160 OpenAI, AI safety (n 13)

161 See, Kali Hays and Alistair Barr, 'AI is Killing the Grand Bargain at the Heart of the Web. We're in a Different World' (*Business Insider*, 2 January 2024) <<https://www.businessinsider.com/ai-killing-web-grand-bargain-2023-8?international=true&r=US&IR=T>> last accessed on 31 May 2024; David Pierce, 'The Text File that Runs the Internet' (*The Verge*, 14 February 2024) <<https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders>> accessed 31 May 2024.

activities carried out for journalistic, academic, artistic, or literary expression, the processing activities conducted by OpenAI on publicly accessible online data to train ChatGPT do not fall within the scope these provisions.<sup>162</sup> Nevertheless, OpenAI may try to find leeway through case law by piggybacking on the *de facto* exception created for search engine operators by the CJEU in its *GC and others* judgment.

In this case, the CJEU was asked to clarify whether the prohibition of processing sensitive personal data, enshrined in Article 9(1) GDPR, also applies to search engine operators having regard to their specific responsibilities, powers, and capabilities.<sup>163</sup> While answering this inquiry, the CJEU first underlined that, apart from its exceptions, this prohibition applies to *all* kinds of processing activities.<sup>164</sup> Therefore, it stated that search engine operators should ensure that their processing activities comply with the respective requirements of the EU data protection framework in the context of their responsibilities, powers, and capabilities.<sup>165</sup> The CJEU nevertheless stated that the specific features of the processing activities carried out by these operators may affect the extent of their responsibilities and obligations attached to sensitive data processing.<sup>166</sup> In this regard, the CJEU clarified that search engine operators are responsible for referencing webpages and specifically for displaying the links to those webpages in turn of a search carried out based on an individual's name and are not responsible for the personal data in question to appear on those webpages.<sup>167</sup> Accordingly, it was decided that the prohibition on sensitive data processing and its restrictions apply to search engine operators only because of this referencing and thus via a verification based on de-referencing requests<sup>168</sup> of data subjects.<sup>169</sup>

In this case, the referring court was further inquiring whether search engine operators could refuse de-referencing requests if it appears that the processing in question is covered by the exceptions enshrined in Article 9(2)(a) or (e) GDPR.<sup>170</sup> Regarding the first exception, Article 9(2)(a) GDPR, the CJEU acknowledged that it would not be practically possible for search engine operators to obtain the explicit consent of data

subjects before the referencing took place and a de-referencing request should in any way be considered as the withdrawal of the data subject's consent.<sup>171</sup> Moving on to Article 9(2)(e) GDPR, the CJEU first underlined that this exception also applies to search engine operators, allowing them to rely on this legal basis to process the sensitive data and accordingly deny de-referencing requests if the processing activity in question relates to data that are manifestly made public by the data subject and complies with other obligations, such as data protection principles, to be considered lawful, unless where the data subject is still entitled to that de-referencing.<sup>172</sup> It should be noted here that the CJEU did not further explain in this case what conditions personal data must meet to be considered manifestly made public by the data subject.

Although the referring court inquired only about these two exceptions, the CJEU went further and underlined that Article 9(2)(g) GDPR, which allows sensitive data to be processed where it is necessary for reasons of substantial public interest, based on Union or Member State law that is proportionate to the aim pursued, respects the essence of the right to data protection, and provides for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject, could also be relied on by search engine operators.<sup>173</sup> In this context, the CJEU required search engine operators to examine whether the processing in question falls under Article 9(2)(g) GDPR by checking whether that processing appears to be strictly necessary for the enjoyment of the right of freedom of information of the Internet users and complies with the conditions set forth therein while answering to de-referencing requests related to it.<sup>174</sup>

In sum, with its *GC and others* judgment, the CJEU allowed the verification of the lawfulness of sensitive data indexing by search engine operators to be conducted *ex post* upon a de-referencing request, like a 'notice and takedown' procedure, due to the specific features of this processing activity and having regards to the responsibilities, powers, and capabilities of search engine operators.<sup>175</sup> In other words, the CJEU 'de facto created an exception which is (...) only applicable to search engine

162 See, C-131/12, *Google Spain SL and Google Inc.* (n 38) para 85.

163 *Ibid* para 31.

164 *Ibid* para 42.

165 *Ibid* para 43.

166 *Ibid* para 45.

167 *Ibid* para 46.

168 De-referencing requests correspond to cases where individuals ask search engine operators to remove certain web pages from the list of results displayed to their users when a search is conducted using their names.

169 C-136/17, *GC, AF, BH, ED* (n 95) para 47.

170 *Ibid* para 49.

171 *Ibid* para 62.

172 *Ibid* paras 63–65.

173 *Ibid* para 61.

174 *Ibid* paras 66–68.

175 Jure Globocnik, 'The Right to be Forgotten is Taking Shape: CJEU Judgments in *GC and Others* (C-136/17) and *Google v CNIL* (C-507/17)' (2020) 69 *GRUR International* 380, 380; Yuliya Miadzvetskaya and Geert Van Calster, 'Google at the Kirchberg Dock. On Delisting Requests, and on the Territorial Reach of the EU's GDPR' (2020) 6 *European Data Protection Law Review* 143, 146.

operators'.<sup>176</sup> Such an approach derived from the practical impossibility of *ex-ante* compliance checks by search engine operators regarding whether the indexed webpages contain sensitive personal data and, if so, which legal basis could be used to process them.<sup>177</sup> However, by doing so, the CJEU significantly lowered the high(er) protection afforded to sensitive personal data by effectively letting search engine operators circumvent the prohibition imposed on their processing. It is, nevertheless, quite understandable that the CJEU tried to provide an alternative and pragmatic solution to the problem it had to deal with, the reconciliation of the right to protection of personal data and the freedom of expression and information in the context of indexing of webpages by search engines, considering the severity of the potential implications of giving a contrasting interpretation, such as obliging search engine operators to check *ex-ante* and systematically whether their search results display websites processing sensitive data, which was considered 'neither possible nor desirable' by Advocate General Szpunar in his opinion.<sup>178</sup>

It must be underlined here that, in line with the relevant European Court of Human Rights case law,<sup>179</sup> the CJEU has already acknowledged that the Internet has become one of the means of exercising the right to freedom of expression and information of the public.<sup>180</sup> In this regard, websites, particularly online content-sharing platforms, were considered to play an essential role in, among others, facilitating the dissemination of information.<sup>181</sup> It was argued that by recognizing search engine operators as 'actors diffusing and making available the information generated by others' in this judgment, the CJEU kept pace with the changing trends regarding the online dissemination of information.<sup>182</sup> In other words, the CJEU considered the activities of search engines in 'finding information published or placed on the internet by third parties, indexing it automatically, storing it temporarily and, finally, making it

available to internet users according to a particular order of preference'<sup>183</sup> are of crucial importance for the enjoyment of the freedom of information of the Internet users. Therefore, this *de facto* exception finds its root in Recital 4 GDPR, which was explicitly referred to in the *GC and others* judgment and states that 'The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality.'

However, some authors have voiced their concerns on this *de facto* exception by underlining that search engine operators could (continue to) process sensitive personal data without relying on a (valid) legal basis in the absence of de-referencing requests.<sup>184</sup> Most importantly, it was speculated that data controllers who also process publicly accessible online data with similar methods, for example crawling, could try to piggyback on this exception created for search engine operators, which may put the effective and full protection of data subjects and legal certainty principle at stake.<sup>185</sup>

It may be that OpenAI will become one of those who try to piggyback on this *de facto* exception. In fact, such a scenario may not be that far-fetched, considering that OpenAI's updated privacy policy refers to the legitimate interest of the broader society in their processing activities related to training and improving ChatGPT.<sup>186</sup> Furthermore, when individuals who do not have an OpenAI account want to 'Make a Privacy Request' and ask OpenAI to remove their personal data, OpenAI states that:

Under certain privacy or data protection laws, such as the GDPR, you may have the right to object to the processing or request removal of your personal data from OpenAI's models or products. (...) OpenAI will verify and consider your request, balancing privacy and data protection rights with public interests like access to information, in accordance with applicable law.<sup>187</sup>

176 Globocnik (n 175) 383. See also, Silvia De Conca, 'GC et al v CNIL: Balancing the Right to Be Forgotten with the Freedom of Information, the Duties of a Search Engine Operator' (2019) 5 European Data Protection Law Review 561, 565.

177 De Conca *ibid* 565.

178 Opinion of Advocate General Szpunar in *G.C., A.F., B.H., E.D. v Commission nationale de l'informatique et des libertés (CNIL)*, Case C-136/17, ECLI:EU:C:2019:14, paras 44–49.

179 See eg, *Fredrik Neij and Peter Sunde Kolmisoppi v Sweden*, App no 40397/12 (European Court of Human Rights, 19 February 2013); *Cengiz and Others v Turkey*, App nos 48226/10 and 14027/11 (European Court of Human Rights, 1 March 2016) paras 49–52.

180 Freedom The right to freedom of expression and information in the EU jurisprudence is enshrined in art 11 of the Charter of the Fundamental Rights of the European Union: '(1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. (2) The freedom and pluralism of the media shall be respected.'

181 See, for instance, *Republic of Poland v European Parliament and Council of the European Union*, Case C-401/19, [2022] ECLI:EU:C:2022:297, paras 45–46.

182 De Conca (n 176) 565–66; Globocnik (n 175) 383.

183 C-136/17, GC, AF, BH, ED (n 95) para 35.

184 Globocnik (n 175) 383.

185 *ibid*; De Conca (n 176) 565; Shay Buckley, 'Defamation Online—Defamation, Intermediary Liability and the Threat of Data Protection Law' (2020) 19 *Hibernian Law Journal* 82, 106; Orla Lynskey, 'Delivering Data Protection: The Next Chapter' (2020) 21 *German Law Journal* 80, 82. See further, Autoriteit Persoonsgegevens, *Scraping* (n 68) 18.

186 OpenAI, *Europe privacy policy* (n 45).

187 'OpenAI Privacy Request Portal' (OpenAI, 12 January 2024) <<https://privacy.openai.com/policies>> accessed 2 June 2024. In order to see this information, data subjects need to 'Submit privacy requests through this portal by clicking the "Make a Privacy Request" button on the top right of this page', select the 'I don't have an OpenAI account' option on the popping-up banner, click on the 'OpenAI Personal Data Removal Request' option, and verify their identity by clicking on a link sent to their email address by OpenAI.

This procedure, however, highly resembles the ‘de-referencing procedure’ established in the *GC and others* judgment. Moreover, in its recently published report on the work undertaken by the ChatGPT Taskforce, the EDPB, regarding the lawfulness of OpenAI’s processing activities to train ChatGPT with publicly accessible online data, stated that: ‘In the present context, where large amounts of personal data are collected via web scraping, a case-by-case examination of each data set is hardly possible.’<sup>188</sup> All these indicate that OpenAI may soon claim that the *de facto* exception created for search engine operators also applies to its processing activities when training ChatGPT with publicly accessible online data.

However, such a claim may not be easily accepted since there are certain differences between the responsibilities, powers, and capabilities of search engine operators and developers of LLMs, as well as the specific features of processing activities conducted for indexing webpages and training ChatGPT with publicly accessible online data. Before mentioning such differences, nevertheless, it should be acknowledged that there are indeed some similarities between search engines and LLMs. For example, while search engines use methods such as crawling to find and index online information, similar methods are used to train LLMs with publicly accessible online data. Most importantly, both tools could be used to access the information.

Yet, there are also fundamental differences between search engines and LLMs. First and foremost, search engines, in principle, act solely as intermediaries between users and the source of information, meaning that search engines, in principle, do not make any alterations to the indexed information. In other words, search engines allow users to reach the ‘unprocessed’ information. On the other hand, LLMs further process publicly accessible online data to, among others, learn connections between words to generate their answers by predicting the words most likely to follow each other. Hence, they present their users with ‘processed’ information, which, as explained above, sometimes appears inaccurate. Therefore, it can be argued that LLMs go beyond being solely an intermediary between the source of information and the user since there is a (possible) alteration of the information throughout the process. Secondly, while there are certain legal and technical ways to prevent search engines from indexing particular information, the same cannot be easily said for LLMs, at

least for now. In connection with this point, thirdly, it should be underlined that search engine results are more accessible and less variable compared to answers provided by LLMs. These differences, however, hold crucial importance when determining whether LLMs could piggyback on the *de facto* exception created for search engines. This is because, with this *de facto* exception, the CJEU effectively obliged data subjects to take an active role in protecting their interests by sending de-referencing requests to search engine operators.<sup>189</sup> While it may be practically possible for data subjects to check whether and which of their personal data are indexed by search engines, the same cannot be said for answers provided by LLMs, given the accessibility issues, e.g., paid services, and, more importantly, the randomness and variety of the answers generated by LLMs. In other words, data subjects cannot easily check whether and how LLMs are using their personal data and intervene where necessary as easily as they can do so for search engines, which may adversely affect their right to protection of personal data.<sup>190</sup>

Given these differences, it seems unlikely for OpenAI to piggyback on this *de facto* exception. In fact, the fact that OpenAI’s processing activities to train ChatGPT with publicly accessible online data has already started to be scrutinized by several DPAs, including the establishment of the dedicated task force on ChatGPT by EDPB, decreases the possibility of OpenAI relying on such a ‘laissez-faire’ interpretation. Even in the scenario where OpenAI relies on this *de facto* exception, it should be underlined that OpenAI should demonstrate the valid legal basis *ex-post*, as otherwise, its processing activities would be considered unlawful.<sup>191</sup> In other words, if we apply the conclusions of the *GC and others* judgment in this context, when data subjects send ‘de-referencing’ requests to OpenAI, it should evaluate whether the personal data in question falls under the exceptions covered by Article 9(2)(a), (e), or (g) GDPR. As explained above, while it is not practically possible for OpenAI to obtain the explicit consent of individuals whose publicly accessible online data are used for training ChatGPT, the ‘de-referencing’ request should be in any way interpreted as the withdrawal of the data subject’s consent. Furthermore, the analysis provided in earlier sections demonstrated that, in some instances, it would be unlikely for OpenAI to meet the conditions of Article 9(2)(e) GDPR when processing publicly accessible data to train ChatGPT. This, however, leaves Article 9(2)

188 EDPB, ChatGPT (n 29) para 19, page 7.

189 Globocnik (n 175) 383.

190 See further, EDPS, TechSonar (n 71) 8; Lilian Edwards and others, ‘Private Ordering and Generative AI: What Can We Learn from Model

Terms and Conditions?’ CREATE Working Paper 2024/05, 17–19. For a practical example to this concern, see, Verma and Oremus (n 65).

191 See, European Data Protection Board, ‘Guidelines 5/2019 on the criteria of the Right to be Forgotten in the search engines cases under the GDPR (part 1)’ V 2.0 (7 July 2020), para 36, page 10.

(g) GDPR as the most feasible exception that OpenAI could rely on in such cases. This means that if OpenAI piggybacks on this *de facto* exception and receives a “de-referencing” request from a data subject, the most feasible option for OpenAI to deny such a request would be to demonstrate that the processing in question is strictly necessary for the enjoyment of the freedom of information of the Internet users and compliance with the conditions outlined in Article 9(2)(g) GDPR.

However, this may not be an easy task for OpenAI. This is because, to demonstrate compliance with Article 9(2)(g) GDPR, OpenAI should attest to the substantial public interest pursued with the processing in question. Indeed, in the *GC and others* judgment, the CJEU acknowledged the enjoyment of the right of freedom of information of Internet users as such.<sup>192</sup> However, in the following *TU, RE v Google LLC*<sup>193</sup> judgment, the CJEU explicitly stated that the right to inform and be informed could not include the right to disseminate and have access to inaccurate information.<sup>194</sup> In this context, it should primarily be underlined that OpenAI's privacy policy provides an explicit note about accuracy, which states that:

Services like ChatGPT generate responses by reading a user's request and, in response, predicting the words most likely to appear next. In some cases, the words most likely to appear next may not be the most factually accurate. For this reason, you should not rely on the factual accuracy of output from our models. (...) Given the technical complexity of how our models work, we may not be able to correct the inaccuracy in every instance.<sup>195</sup>

Moreover, the Italian DPA also used this inaccuracy issue to substantiate its ban on ChatGPT in early 2023.<sup>196</sup> Accordingly, it is open to debate whether and to what extent OpenAI could claim that its processing activities conducted on publicly accessible online data to train ChatGPT are strictly necessary for the enjoyment of the freedom of information of Internet users due to this unsolved inaccuracy problem.

Even in a case where these processing activities are considered to be strictly necessary for such purposes, OpenAI would have to demonstrate compliance with other conditions enshrined in Article 9(2)(g) GDPR,

which requires the processing activities to be based on a Union or Member State law which is proportionate to the aims pursued, respect the essence of the right to data protection, and provide for suitable and specific measures to safeguard the fundamental rights and interests of the data subjects. In the *GC and others* case, partly due to the framing of the questions referred to, the CJEU did not engage further with these conditions, leaving the discussion on whether there is any such law that search engines could rely on somewhat ‘glossed over’.<sup>197</sup> It should be noted here that, as per Article 85(1) GDPR, Member States are obliged to reconcile by law the right to the protection of personal data with the right to freedom of expression and information, and it is further argued that these laws should be formulated per the conditions outlined in Article 9 (2)(g) GDPR when such reconciliation is related to the processing of sensitive data.<sup>198</sup> Yet, Erdos reported in 2021 that only a few Member States passed specific laws in their legislation as per Article 85(1) GDPR, while the adequateness of these laws appeared as another concern.<sup>199</sup> The absence of such specific laws, however, could put the validity of reliance on Article 9 (2)(g) GDPR by OpenAI at stake. Although Erdos opines that Article 11 of the Charter of the Fundamental Rights of the European Union and constitutional provisions in Member State laws could be used in the absence of such specific laws, he nevertheless underlines the need for appropriately balanced and safeguarded processing activities in such cases.<sup>200</sup> In light of the concerns mentioned in earlier sections of this article, it is open to debate whether and to what extent processing activities conducted on publicly accessible online data to train LLMs comply with these requirements in their current form. Furthermore, as the CJEU underlined in the *Google v CNIL*<sup>201</sup> judgment, balancing these competing rights would differ across Member States, which may cause fragmentation.<sup>202</sup> Besides, some cases have been reported in which DPAs declared sensitive data processing activities based on Article 9(2)(g) GDPR unlawful due to inexistent or inadequate Member State laws.<sup>203</sup> Accordingly, it can be argued that even such specific

192 C-136/17, GC, AF, BH, ED (n 95) para 66. See also, Globocnik (n 175) 383; De Conca (n 176) 565.

193 *TU, RE v Google LLC*, Case C-460/20, [2022] ECLI:EU:C:2022:962.

194 Ibid para 65.

195 OpenAI, Europe Privacy Policy (n 45).

196 Garante (n 21).

197 See, Mark Leiser and Bart Schermer, 'GC & Others vs CNIL and Google: This is a Special Case' (*European Law Blog*, 20 November 2019) <<https://europeanlawblog.eu/2019/11/20/gc-others-vs-cnil-and-google-this-is-a-special-case/>> accessed 2 June 2024.

198 Erdos (n 151) 417-425. See also, Case IMY-2022-1621, Integritetsskyddsmyndigheten [2022] <[https://gdprhub.eu/index.php?title=IMY \(Sweden\) - IMY-2022-1621](https://gdprhub.eu/index.php?title=IMY_(Sweden)_-_IMY-2022-1621)> accessed 2 June 2024.

199 Erdos (n 151) 420. See also, Athena Christofi, Ellen Wauters and Peggy Valcke, 'Smart Cities, Data Protection and the Public Interest Conundrum: What Legal Basis for Smart City Processing?' (2021) 12 *European Journal of Law and Technology* 21.

200 Erdos (n 151) 427–28.

201 *Google LLC v Commission nationale de l'informatique et des libertés (CNIL)*, Case C-507/1, [2019] ECLI:EU:C:2019:772.

202 Ibid para 67.

Member State laws could only provide unreliable legal grounds for OpenAI to process publicly accessible online data to train ChatGPT, as their inconsistency and quality are liable to pose practical and legal challenges.

In sum, even if OpenAI relies on the *de facto* exception granted to search engine operators in the *GC and others* judgment, whether and to what extent it could demonstrate the valid legal basis for the processed data appears questionable, which could prevent this possibility in the first place. However, this outcome could jeopardize an emerging and promising business model, which may oblige the enforcers of the EU data protection framework to look for alternatives, as the CJEU did in the *GC and others* judgment. Such a solution, on the other hand, could put the legal certainty principle and the effective and full protection of the right to data protection at stake.

## It takes a village

As illustrated through the case of ChatGPT, one of the biggest battles in EU data protection law will soon be finding, if it exists, the proper legal basis for training LLMs with publicly accessible online data. However, even when such a legal basis is found, each stakeholder has certain obligations to prevent the abuse and misuse of personal data that is used for training these models and to strike a proper balance between competing interests without hindering innovation.

First, the developers of LLMs should adhere to all data protection principles. In this regard, respecting the transparency obligations is of crucial importance. Accordingly, these developers should provide the required information about their processing activities to all data subjects indiscriminately. While these developers could enjoy several exemptions to this obligation, such as the one enshrined in Article 14(5)(b) GDPR, by claiming that providing such information proves impossible or requires disproportionate effort given the vast number of data subjects,<sup>203</sup> it should be underlined that they would still be obliged to take certain measures, for example making the information about their processing activities publicly available, to protect the fundamental

rights and freedoms of these individuals. Relating to this point, it should be recalled that the Italian DPA requested OpenAI to present a notice on its website describing its processing activities to its users and beyond.<sup>205</sup> Moreover, OpenAI was asked to run a non-marketing information campaign on Italian mass media to inform Italians about its processing activities.<sup>206</sup> Similarly, the developers of LLMs should come up with effective strategies to inform the public about their processing activities.

Most importantly, compliance with the data protection by design and by default principle throughout the lifecycle of LLMs should be ensured. As explained in earlier sections, the vast amount of data processed to train these models has already led to serious concerns and legal disputes. Therefore, it appears, in fact, as a necessity and a priority for developers of these models to comply with this principle to continue their operations. In fact, such efforts would also lead them to comply with data minimization and purpose limitation principles, which oblige only the required and limited amount of personal data to be processed for specific purposes outlined at the beginning of processing activities.<sup>207</sup> Eventually, like Google tailored its practices regarding Street View as a result of the privacy concerns voiced against it by, for example, introducing options to blur, among others, houses, bodies, and license plates,<sup>208</sup> developers of LLMs should come up with technical strategies to only process personal data that comply with the data protection requirements and to ‘unlearn’ those that contradict these requirements. Lastly, OpenAI mentions that ‘We have also completed a data protection impact assessment to help ensure we are collecting and using [the personal information that is included in training information] legally and responsibly.’<sup>209</sup> Although not mandatory, it may be useful for OpenAI and other developers to disclose their data protection impact assessment reports or summaries to ensure data subjects understand how their personal data are processed and assess how the related risks are mitigated.

While they have limited abilities to prevent scraping activities,<sup>210</sup> operators of websites that are used for

203 Integritetsskyddsmyndigheten (n 198); Case 0098/2022, Agencia Española de Protección de Datos [2023] <[https://gdprhub.eu/index.php?title=AEPD\\_\(Spain\)\\_-0098/2022](https://gdprhub.eu/index.php?title=AEPD_(Spain)_-0098/2022)> accessed 2 June 2024. See also, Case 6745/163/18, Tietosuojavaltuutetun toimisto [2021] <[https://gdprhub.eu/index.php?title=Tietosuojavaltuutetun\\_toimisto\\_\(Finland\)\\_-6745/163/18](https://gdprhub.eu/index.php?title=Tietosuojavaltuutetun_toimisto_(Finland)_-6745/163/18)> accessed 2 June 2024.

204 GDPR, recital 62. See also, EDPB, ChatGPT (n 29) para 27, page 8.

205 Garante (n 43)

206 Ibid.

207 ‘In reality, these tools can be built with less data and without coercive and secretive data collection processes’. See, Electronic Privacy Information Center (n 75) 24.

208 See, ‘Google-Contributed Street View Imagery Policy’ (Google) <<https://www.google.com/streetview/policy/>> accessed 1 November 2023.

209 OpenAI, ChatGPT (n 25)

210 Hays and Barr (n 161); Thilo Gottschalk, ‘The Data-Laundromat? Public-Private-Partnerships and Publicly Available Data in the Area of Law Enforcement’ (2020) 6 European Data Protection Law Review 21, 27.

extracting data to train LLMs also bear certain obligations to adhere to data protection principles, specifically data protection by design and by default.<sup>211</sup> In this regard, they, especially those processing sensitive data, should inform their data subjects effectively about attached risks, for example unrelated further processing, when disclosing their personal data. To mitigate such risks, they might consider also implementing scraping-preventive measures.<sup>212</sup> In fact, it is reported that, at the time of the writing of this article, around 25 per cent of the Top 1000 websites have already blocked GPTBot, including websites such as Amazon, Quara, and Indeed.<sup>213</sup>

Data subjects may also take certain measures to protect their rights and interests. For example, the 'Joint statement on data scraping and the protection of privacy' issued by 12 DPAs from around the world states that data subjects should carefully read the information provided by controllers, especially regarding disclosure policies, to make informed decisions by having a long-term vision about making their personal data accessible online. Furthermore, data subjects are advised to be mindful of the types and amount of personal data they share online to mitigate the risks if their personal data are further processed for unrelated or malicious purposes.<sup>214</sup>

Lastly, regulators and DPAs should ensure legal clarity for all stakeholders involved. They should, for example, oblige developers of these models to provide more transparency regarding their operations to understand whether and to what extent they respect data protection principles and obligations. In this regard, the EDPB's recent efforts in creating a special task force on ChatGPT to foster cooperation and exchange information on possible enforcement actions conducted by DPAs should be appreciated.<sup>215</sup> Through such initiatives, guidance should be provided on currently unclear matters, such as the proper legal basis to process personal data for training these models. Likewise, with the increasing awareness about the potential implications of further (unrelated) uses of publicly accessible online data, stricter scrutiny should be ensured for data controllers disclosing these data to the public, especially those who process sensitive personal data.

## Conclusions

Since its launch in November 2022, ChatGPT has been the hottest topic of data protection discussions and beyond. While too much is expected from ChatGPT and LLMs in general, serious concerns have been raised regarding their risks, such as privacy and data protection concerns, copyright infringements, spreading misinformation, and cheating in examinations. Therefore, finding the perfect balance when regulating LLMs to allow and foster innovation without hindering fundamental rights and freedoms holds crucial importance.

To that end, this article addressed one of the primary data protection concerns by taking ChatGPT as a case study: whether there is a valid legal basis for processing publicly accessible online data to train LLMs. Although this question has been examined through Article 6 GDPR so far, given the recent judgments of the CJEU that shed light on what constitutes sensitive data and which regime applies when sensitive and non-sensitive data are processed *en bloc*, it appeared as Article 9 GDPR should be visited instead to answer this question. Accordingly, the article evaluated whether OpenAI could rely on Article 9(2)(a) GDPR, which allows sensitive data to be processed based on explicit consent of the data subject, or Article 9(2)(e) GDPR, which allows personal data that is manifestly made public by the data subject to be processed, for training ChatGPT with publicly accessible online data. As it was argued that it would be practically very hard, if not impossible, for OpenAI to obtain the explicit consent of probably millions of individuals subjected to these processing activities, Article 9(2)(e) GDPR appeared as the most suitable exception for OpenAI.

When the test developed by the CJEU in the *Meta v Bundeskartellamt* case applied to the processing activities in question, however, it was revealed that only a limited amount of personal data that is publicly accessible online could actually meet the criteria to be considered as 'manifestly made public by the data subject'. Hence, the article called for further clarification to understand whether OpenAI processes only this limited amount of personal data that complies with those requirements to train ChatGPT and underlined that the lawfulness of these processing activities would be at

211 In fact, the responsibilities of these actors have long been called upon in terms of protecting the privacy of their users, see eg, Eleni Kosta and others, 'Data Protection Issues Pertaining to Social Networking under EU Law' (2010) 4 *Transforming Government: People, Process and Policy* 193.

212 See, OpenAI, GPTBot (n 14); Hays and Barr (n 161); 'Intelligenza Artificiale: Dal Garante Privacy le Indicazioni per Difendere i dati Personali dal Web Scraping' (*Garante per la protezione dei dati personali*, 30 May 2024) <<https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10019984#english>> accessed 31 May 2024.

213 Jonathan Gillham, 'Websites that have Blocked OpenAI's GPTBot CCBot Anthropic Google Extended - 1000 Website Study' (*Originality.ai*) <<https://originality.ai/ai-bot-blocking>> accessed 1 November 2023.

214 'Joint Statement on Data Scraping and the Protection of Privacy' (*Information Commissioner's Office*, 24 August 2023) paras 18–20, pages 4–5 <<https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf>> accessed 31 October 2023.

215 EDPB, Task Force (n 23).

stake until such clarity is ensured, which is a concern that also applies to other developers who follow the same approach to train their models.

At this point, the article speculated about a possible way forward and explored whether OpenAI could piggyback on the *de facto* exception created in the CJEU jurisprudence for search engine operators regarding sensitive data processing. Although there are already some indications that OpenAI may soon follow this path, the article argued that it may not be a very strong case given the differences between the responsibilities, powers, and capabilities of search engine operators and developers of LLMs, as well as the specific features of processing activities conducted for indexing webpages and training ChatGPT with publicly accessible online data. Furthermore, some legal uncertainties that might compromise the lawfulness of processing activities conducted on publicly accessible online data to train ChatGPT were underlined even where OpenAI relies on this *de facto* exception. Lastly, the article provided an overview of the obligations of all stakeholders involved to ensure that LLMs are developed responsibly until and even after the proper legal basis to train them with publicly accessible online data is found.

Consequently, it is safe to assume that the search for the proper legal basis for training LLMs with publicly accessible online data will be ‘the next big thing’ in the EU data protection framework. While there are indeed other data protection-related issues attached to these models, naturally, addressing all these concerns is beyond the aims and limitations of this article. Nevertheless, it should be underlined here that OpenAI states that ‘we believe that society must have time to update and adjust to increasingly capable AI, and that everyone who is affected by this technology should have a significant say in how AI develops further.’<sup>216</sup> Indeed, LLMs could fulfil their promises without hindering our fundamental rights and freedoms only through cooperation between all stakeholders. Therefore, at the climax of the EU’s efforts to regulate artificial intelligence, I hope the outcomes of this research will trigger a broader social, cultural, and political dialogue to understand what constitutes ‘public’ in the digital sphere today and whether and to what extent could this sphere be exploited for commercial purposes as it is not just a legal question.

<https://doi.org/10.1093/idpl/ipae013>

216. OpenAI, AI Safety (n 13).

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

International Data Privacy Law, 2024, 00, 1–26

<https://doi.org/10.1093/idpl/ipae013>

Article