CDT Europe’s Response to the CNIL Consultation on the development of AI systems

Factsheet 8 - Relying on the legal basis of legitimate interests to use personal data to develop an AI system

CDT welcomes CNIL’s re-statement of the three cumulative conditions that must be complied with in order to rely on the legitimate interests legal basis, namely i) the interest must be legitimate, ii) the processing must be necessary, and iii) the objective pursued must not threaten the rights and interests of data subjects.

CDT acknowledges that some interests could be a priori legitimate. However, the factsheet should clarify that the legitimate interest cannot be used as a legal basis to process sensitive data, and narrowly define the interests which would constitute an “a priori” legitimate interest. For example, the draft sheet states that developing new systems and functionalities for users of a service would be a priori a legitimate interest, without setting meaningful guardrails that would prevent this particular interest from being exploited. Without those limits, developers of foundation models would have wide latitude to inappropriately rely on the presumptive lawfulness of the interest identified as “a priori” legitimate. One need only look to the United States to see that companies can and will take advantage of such a broad legitimate interest, as companies like Google and Meta changed their practices and in some cases updated their privacy policies to allow for widespread use of data for developing and training AI products.[1] The CNIL should therefore explore using different language to ensure that the development of new systems and functionalities is only an a priori legitimate interest if tied to users’ reasonable expectations.

CDT notes that the factsheet is silent on the aspect of informing data subjects. We recall that informing data subjects is an essential requirement to prove the necessity of the data processing as part of the legitimate interests assessment, as noted by the CJEU[2] - and therefore this aspect should be addressed in factsheet 8, irrespective of the fact that the topic is generally covered in Sheet 9. It should be reiterated that, as indicated by the EDPB in their ChatGPT’s Taskforce report,[3] data subjects should be “clearly and demonstrably informed” that personal data processed under the legitimate interest legal basis - contained in either prompts, uploads or feedback regarding the output - will be used for training purposes. The factsheet should therefore

[1] Eli Tan, *When the Terms of Service Change to Make Wai for AI Training*, 26 June 2024, The New York Times. Available at: https://www.nytimes.com/2024/06/26/technology/terms-service-ai-training.html

[2] CJEU, *Meta Platforms Inc and Others v Bundeskartellamt*, Case C-252/21, para. 126. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62021CJ0252

[3] European Data Protection Board, Report of the work undertaken by the ChatGPT Taskforce, 23 May 2024. Available at: https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf

emphasise that relying on a legitimate interest legal basis does not exempt developers from notifying data subjects.

The economic interest of the AI system developer alone cannot justify an interference with an individual's rights and interests.[4] In the *Google Spain* case, the CJEU noted in its analysis that, while the company's legitimate interests could not prevail over the individual's rights and interests to have personal data removed from a search engine on that set of facts, the removal of personal data corresponding to that individual could have effects on the legitimate interest of internet users using the company's search engine, and that a fair balance should be sought between the interests of the data subject and those of the interests of users. While the Court noted that the data subject's rights would as a general rule override the interest of others, the balance could depend on the nature of the information in question, its sensitivity, and the interest of the public in having that information. The Court's analysis in this case is a helpful starting point for the CNIL, taking into consideration the fact that public usage of the most popular search engine in the world is likely to be far higher than the usage of any particular AI model, therefore weakening any AI model developer's claim that the legitimate interests relied upon are those of its users.

CDT appreciates the emphasis on data minimisation when training AI models, particularly in light of recent research showing that at least in some contexts, additional data can lead to worse model outcomes.[5] While the "Taking into account data protection when designing the [AI] system" sheet provides some guidance in determining how a developer is to operationalize the data minimisation principle, it is crucial for CNIL to provide clearer guidance to avoid companies engaging in practices that may lead to infringements of the GDPR. For instance, CNIL could consider stating that it is presumptively not "necessary" to collect and store personal data that constitutes sensitive data about individuals indefinitely and for purposes of training an AI system. Further, CNIL could require AI developers to explain why they cannot design an AI system with less personal data, and why identifiable sensitive data is vital to the training of the system.

Factsheet 8.1 - The legal basis of legitimate interests: Focus sheet on open source models

CDT highlights that open models can provide significant benefits to society, including the advancement of competition and research, the protection of civil and human rights, transparency, and safety and security. While open source models pose some risks, recent research highlights the importance of evaluating the risks of open models not in a vacuum, but in comparison to the

[4] CJEU, *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos* (AEPD), Case C-131/12, para.81. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0131

[5] Judy Hanwen Shen et al. "The Data Addition Dilemma", *Proceedings of Machine Learning Research* 252:1-43, 2021. Available at: https://arxiv.org/pdf/2408.04154

risks and benefits from closed models and pre-existing technologies like the internet.[6] In deciding its approach to open source models, CDT calls for the CNIL to focus on the marginal risks that specifically arise from open source models. When a model is released with its training data, independent third parties can better assess the model's capabilities and risks. Researchers have found that while evidence establishing significant marginal risk from open foundation models is generally scarce, in some cases - for example in relation to AI-generated child sexual abuse material (CSAM) and nonconsensual intimate imagery (NCII) - the harms stemming specifically from open foundation models have been better documented.[7]

In addition to assessing open source models by reference to marginal risks, CDT encourages the CNIL to take steps to encourage genuine transparency on training data, beyond what is currently recognised by the AI Act. Under the Act, only AI systems recognised as high-risk are required to disclose training data,[8] with systems and models falling outside this classification being exempt from the requirement. Disclosing information on training datasets is not a pre-requisite to benefit from the "open source" label. In fact, the AI Act considers that general-purpose AI models released under free and open-source licences should be considered to ensure high levels of transparency if parameters, weights, information on the model architecture, and information of model usage are made publicly available,[9] without any mention being made of transparency of datasets. Several commentators have noted this shortcoming,[10] and similarly flagged the risk of licenses being used as a tool for "open-washing". The CNIL could therefore be well-placed to demand heightened transparency on training data.

Factsheet 8.2 - The legal basis of legitimate interests: Focus sheet on measures to implement in case of data collection by web scraping

CDT welcomes CNIL's direct assessment of the suitability of legitimate interests as a legal basis for web scraping under the General Data Protection Regulation. Web scraping poses serious challenges from a data protection perspective, not least because the data subject is rarely aware that their personal data is or will be scraped. Recital 47 of the GDPR notably states that the

[6] Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," Center for Research on Foundation Models (CRFM), Stanford University, February 2024.

[7] Stanford Institute on Human-Centered Artificial Intelligence, *Considerations for governing open foundation models*, Issue Brief, December 2023. Available at: https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

[8] EU Artificial Intelligence Act, Article 11(1), and Annex IV.

[9] EU Artificial Intelligence Act, Recital 102.

[10] Open Future, AI Act fails to set meaningful dataset transparency standards for open source AI. Available at: https://openfuture.eu/blog/ai-act-fails-to-set-meaningful-dataset-transparency-standards-for-open-source-ai/; Andreas Liesenfeld and Mark Dingemanse, Rethinking Open Source Generative AI: Open Washing at the EU AI Act, *FAccT '24: Proc. 2024 ACM Conf. on Fairness, Accountability, and Transparency* 1774–1787 (ACM, 2024). Available at: https://dl.acm.org/doi/pdf/10.1145/3630106.3659005

assessment of legitimate interest shall take into consideration the reasonable expectations of data subjects based on their relationship with the controller, 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. In particular, the recital underscores that the interests and fundamental rights of the data subject could override the interest of the data controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing.

Data subjects rarely have awareness of, much less a direct relationship with, the entities conducting web scraping. Instead, data subjects interact with the websites holding their data, which themselves hold little power to prevent any data on their websites from being scraped. There are known limits to the extent that webpages can protect their data from being scraped, as many web crawlers disregard the Terms of Service of a particular website, and can circumvent the instructions in a website's Robots Exclusion Protocol (known as robots.txt file). Consequently, the argument that there could be a reasonable expectation for personal data to be scraped is implausible in most cases. A recent decision from the Dutch DPA in connection with Clearview found that there was no relationship whatsoever between Clearview and the data subjects whose personal data had been included in the Clearview database.[11] Noting the absence of notification to individuals either before or after the scraping, combined with the public inaccessibility of Clearview's database, the DPA held that there was no individual reasonable expectation that Clearview could scrape the data, regardless of the fact that the data scraped was public.[12] CDT agrees.

In light of the potential intrusiveness and generally indiscriminate nature of web scraping, CDT makes two recommendations to the CNIL. Firstly, the CNIL should distinguish web scraping from other activities involved in the development of an AI system, and explicitly narrow the scenarios in which legitimate interests could be relied upon as valid legal basis specifically for data collection via web scraping. The fact sheet's approach does not distinguish between the different purposes for which web scraping may be conducted on a "legitimate interests" legal basis. In CDT's view, the range of interests which would be presumed legitimate in order to lawfully conduct web scraping should be narrower than the range of interests identified in fact sheet 8 as a priori legitimate in connection with the development of an AI system. For example, improving a product or a service to increase its performance, or offering the service of a conversational agent to assist users - both identified in fact sheet 8 as interests that are a priori legitimate - should not be considered a priori legitimate for the purposes of web scraping. Conversely, scientific or academic research could in some circumstances be a priori legitimate. These and the preceding interests should be sufficiently differentiated, with additional wording to prevent non-academic secondary uses of academic databases built off scraped material.

[11] See Autoriteit Persoonsgegevens, Decision dated 16 May 2024, para. 117. Available at: https://www.autoriteitpersoonsgegevens.nl/en/current/dutch-dpa-imposes-a-fine-on-clearview-because-of-illegal-data-collection-for-facial-recognition

[12] Autoriteit Persoonsgegevens, Decision dated 16 May 2024, para. 118.

Secondly, in the narrow cases where legitimate interests could be used as a valid legal basis for web scraping, CNIL should broaden the safeguards categorised as "mandatory". Many of the safeguards categorised by the CNIL as "additional" should be instead considered mandatory, such as the obligation to exclude collection from websites which clearly object to scraping, disseminate widely information about data collection and data subjects' rights, and applying anonymisation or pseudonymisation processes immediately after data collection.

Factsheet 9 - Informing data subjects

CDT supports the CNIL's intention to provide clarity on the GDPR obligation to inform data subjects of processing activities in light of the challenges AI presents.

The development of AI involves multiple layers which compound the dearth of information available to individuals as to the use of their data. Firstly, individuals are generally not informed whether their data has been scraped into a training dataset or corpus. Secondly, it is increasingly difficult to find publicly available information on the corpora used by AI developers, who have gradually come to release few details about their training data: for example, Meta has made less information available for its latest model Llama 3 than it did for Llama 2, and OpenAI decided not to disclose information on the training datasets used for GPT-4 citing competitiveness concerns. Thirdly, the data does not live in a single place, as corpora that are non-proprietary may be copied or downloaded by researchers and developers, making it difficult for individuals to exercise control over their data once it lives in a dataset. The obligations laid out in GDPR apply regardless of the complexity of the AI development ecosystem, and must be observed to facilitate data subjects' exercise of their rights.

There are already initiatives that allow online users to find out if some of their data have been used to train AI, e.g. Spawning AI's "Have I Been Trained" tool, which allows individuals to verify if their art or photos have been used for training purposes. While these initiatives are positive, GDPR is clear that controllers have an obligation to notify the individual of the processing activities at the time the data is collected, including any legitimate interests relied upon. The onus should not be on the individual to check if their data has been used - rather, GDPR requires the controller to let the individuals know.

CNIL should further clarify the obligations around the rights of rectification and erasure as it relates to AI training datasets. An individual may seek rectification or erasure of their data from a particular AI developer, but that does not mean that the AI model will suddenly be correct or forget that data. AI systems are trained on data at certain points in time, and then that AI model remains "trained" on that data until it is retrained. Individuals will not understand that, and may find themselves harmed even after they request rectification or erasure. CNIL should require AI developers to explain this issue so individuals are not confused about the time delay.

CDT welcomes the CNIL's upholding of the obligation to notify individuals, regardless of whether the data has been directly or indirectly collected. We similarly support the CNIL's support of a latency period before the data collected is actually used for training, but we would further suggest that this period should be no shorter than one month.

On the accessibility of information, CDT would additionally suggest that any notices are created in all languages the company does business in, and should be accessible for people with disabilities.

Factsheet 10 - Respecting and facilitating the exercise of data subjects' rights

As a starting point, CDT encourages the CNIL to anticipate likely objections from AI developers that personal data is difficult to locate, and therefore makes difficult the exercise of data subject rights. When training data encompasses personal data, entities behind the AI system or model should not be allowed to assert that they are unable to specify what personal data was contained within the training data. As asserted by the ChatGPT task force, technical impossibility cannot be invoked to justify non-compliance with GDPR obligations.[13]

Where a model developer collects more data from an individual for purposes of identification of that individual, that data may only be used for that purpose and should be properly disposed of after that purpose is fulfilled. Model developers should not be able to use this exemption for data collection as a loophole.

CNIL should push for periodic model retraining to accommodate individual requests for erasure and rectification. As described above, there is a delay between an individual seeking a request and such time as the request can be honored in the outputs of an AI system. The sooner the AI system respects the requests of the individual, the better.

CDT welcomes the CNIL's comprehensive approach to scope of the right to data access, recalling that the right of access under GDPR is, as previously underscored by the European Data Protection Board, without any general reservation to proportionality with regard to the efforts the controller has to take to comply with the data subject's request.

CDT welcomes the CNIL's suggestion that models should be retrained periodically to "action" data subject requests for opposition, rectification or erasure. However, CDT notes with concern the rationale offered by the CNIL, which considers that this retraining should take place when it is not disproportionate to the rights of the controller, noting that this will depend on the sensitivity of

[13] Studies show that tools already exist which enable the detection of personally identifiable information. See Elazar et al, *What's in My Big Data?,* ICLR 2024 Conference Paper, available at: https://arxiv.org/abs/2310.20707

the data and risks of regurgitation and disclosure to individuals. CDT acknowledges that not all data subject rights are absolute under GDPR, and many allow exemptions. However, none of the existing exemptions include “controllers’s rights”. For example, none of the grounds exempting a controller from the obligation to follow an erasure request include the rights of the controller (GDPR Article 17(3)), nor do they invite an assessment of proportionality. CDT therefore encourages the CNIL to amend the language of the rationale to reflect the requirements of GDPR.