UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Building a Chatbot : Challenges under Copyright and Data Protection Law

Kelli, Aleksei; Tavast, Arvi; Lindén, Krister

Ebers, Martin; Poncibò, Cristina; Zou, Mimi

2022-06-30

http://hdl.handle.net/10138/352386

Kelli, A, Tavast, A & Lindén, K 2022, Building a Chatbot : Challenges under Copyright and Data Protection Law. in M Ebers, C Poncibò & M Zou (eds), Contracting and Contract Law in the Age of Artificial Intelligence. Hart publishing, Oxford, pp. 115-134. https://doi.org/10.5040/9781509950713.ch-007

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

7

Building a Chatbot: Challenges under Copyright and Data Protection Law

Aleksei Kelli, Arvi Tavast, Krister Lindén

I. Introduction

Chatbots have become an integral part of everyday life. A chatbot (conversational agent, dialogue system, virtual assistant) is defined as 'a computer system that operates as an interface between human users and a software application, using spoken or written natural language as the primary means of communication'.[1] Such use of natural language understanding and production makes chatbots one of the most demanding and comprehensive applications of natural language processing (NLP).

The idea behind chatbots is to simulate conversation with another human. In the current stage of technical development, the resulting conversation is not on a human level, and it is obvious for users that they are interacting with an artificial system. Regardless, chatbots are becoming useful in practical applications, especially in clearly defined domains like voice-based search, appointment scheduling or controlling home appliances.

The authors in this chapter focus on copyright and personal data protection challenges relating to building chatbots. The chapter reflects previous research[2] and develops it further. The author's main focus is on building models for chatbots which takes place before the chatbot service is offered.

The authors rely on the EU acquis and also use Estonian law to exemplify legal requirements. The article constitutes interdisciplinary analysis where the authors integrate legal and technological domains.

---

[1] B Galitsky, *Developing Enterprise Chatbots: Learning Linguistic Structures* (Springer, 2019) 13.

[2] A Kelli, A Tavast and K Lindén, 'Vestlusrobotid ja autoriõigus (Chatbots and Copyright)' (2020) 5 *Juridica*; A Kelli, A Tavast, K Lindén, K Vider, R Birštonas, P Labropoulou, I Kull, G Tavits, A Värv, P Stranák and J Hajic, 'The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies' in K Simov and M Eskevich (eds), *Selected Papers from the CLARIN Annual Conference 2019* (53−65). (Linköping University Electronic Press, 2020) 53−65. ep.liu.se/en/conference-article.aspx?series=ecp&issue=172&Article_No=8.

From the legal perspective, chatbots, as such, are computer programs. According to Article 1(1) of the Computer Programs Directive,[3] computer programs are protected by copyright as literary works. Chatbots rely on language models that are copyright-protected databases. A computer program compiles language models (databases) from data snippets. It is not usually possible to extract original data used for the creation of language models. However, the main challenge is the lawful acquisition and use of the training data needed to create language models since the training data could contain copyright-protected works, objects of related rights and personal data.[4]

The adoption and implementation of the DSM Directive[5] have a significant impact on the creation of language models used for chatbots. Articles 3 and 4 of the DSM Directive, which regulate the text and data mining (TDM) exception, are particularly relevant since they establish legal grounds to use training data. Before the DSM Directive, the InfoSoc Directive[6] was the main legal instrument regulating TDM.

When it comes to personal data protection, the adoption of the General Data Protection Regulation[7] (GDPR) has implications for creating chatbots. Personal data protection concerning chatbots is such a relevant and complex issue that the European Data Protection Board (EDPB) has adopted the Guidelines on Virtual Voice Assistants (Guidelines on VVA),[8] systematically analysing the creation and use of chatbots. Scientific literature identifies the following interaction points between generating machine learning models and personal data protection: 1) models cannot be trained from personal data without a specific lawful ground;

---

[3] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Computer Programs Directive) [2009] OJ L111.

[4] The reference to copyrighted content or works also covers objects of related rights. Depending on the context, the terms ‘language data’, ‘training data’ and ‘data’ are used as synonyms. It is presumed that language data contains personal data and copyrighted content.

[5] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (DSM Directive) [2019] OJ L130.

[6] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc Directive) [2001] OJ L167.

[7] 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) [2016] OJ L119.

[8] European Data Protection Board. Guidelines 02/2021 on Virtual Voice Assistants. Version 1.0. Adopted on 9 March 2021 (Guidelines on VVA). edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_022021_virtual_voice_assistants_adopted-public-consultation_en.pdf.

2) data subjects should be informed of the intention to train a model; 3) the data subject has the right to object or withdraw consent; and 4) in case of automated decisions, individuals should have meaningful information about the logic involved.[9]

The first part of our chapter deals with the following technical issues of chatbots: use of machine learning,[10] modality of conversation (spoken or written), location of data processing, length of memory of the chatbot, and purpose of use. The second part concentrates on legal regimes of data used to create a chatbot and deals with the legal basis for the use of data. The authors' focus is on the process of building a chatbot rather than its intended use. It should be pointed out that building and improving a chatbot is a continuous process that does not end with the launch of the service.

II. Technical background of chatbots

A. Use of machine learning

Most chatbots from the first attempts[11] to current systems rely substantially on explicitly programmed rules and manually written responses. For example, if the words 'movies' and 'tonight' can be detected in the user's utterance, the rule could be to output movie listings near the user's location.

In this simple but widespread case, the chatbot has nothing to do with machine learning or related copyright issues. Instead, it is a traditional computer program protected by copyright as a literary work. The main difficulty with such simple chatbots is preparing rules to cover a sufficient variety of possible user inputs with acceptable precision, which can easily become unrealistic unless the domain is narrowly restricted.

Objectives for developing chatbot technology include both improving the quality and reducing the amount of manual labour involved. To get to the legal issues related to this development, let us first briefly describe the working principle of chatbots.

---

[9] M Veale, R Binns and L Edwards, 'Algorithms that remember: model inversion attacks and data protection law' (2018) *Philosophical Transactions of Royal Society A*, 2. doi.org/10.1098/rsta.2018.0083.

[10] See, eg, I Goodfellow, Y Bengio and A Courville, *Deep Learning* (MIT Press, 2016) 96–161. www.deeplearningbook.org/ (31.10.2020).

[11] J Weizenbaum, 'ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine' (1966) 9(1) *Communications of the ACM* 36–45.

Modern chatbots consist of a range of generic NLP components, ie not specific to chatbots, performing the following three main functions:

a) Understanding the user's utterance. Due to variability and under-specification in natural language, this is a non-trivial task. Necessary components may include language detection of which language the user is speaking, text cleaning and error correction, tokenisation identifying sentences and words in an utterance, part-of-speech tagging and syntactic parsing determining the form and function of each word in a sentence, synonym detection in case the user's word choice is slightly different from the system creators', anaphora resolution determining what or whom pronouns in a sentence refer to, named entity recognition whether 'Smith' at the beginning of a sentence is a general noun or a surname, multi-word expression detection recognising that 'give permission' may mean the same as 'allow', etc. Speech recognition is additionally needed for chatbots using the spoken modality.
b) Preparing the response. As already mentioned, chatbots are a type of user interface between a human and an information system. Functions of that information system may range as widely as functions of information systems in general: given a system that can be interacted with by a human, it is at least theoretically conceivable that the interaction takes place in the form of a chatbot.
c) Expressing the response to the user. The task of natural language production is somewhat simpler than natural language understanding and may include the generation of grammatical sentences and the production of natural-sounding text. Speech synthesis or text-to-speech conversion is added in the spoken modality. In multimodal communication, such as a 'talking head' or an animated avatar, an additional task is to generate human-like facial movements to accompany the synthesised speech.

Item b) above is outside the scope of this chapter because information systems, as well as their legal issues, if any, are too heterogeneous to be covered in any detail here. Items a) and c), however, cover all of what is currently considered the domain of NLP.[12]

It is not a coincidence that a talking fridge has previously been used[13] as an example application for discussing legal issues of NLP: a state-of-the-art talking fridge would have to use most if not all technological achievements of the field.

In the current state of machine learning technology, models are primarily trained while developing the chatbot. The complexity of language models and the amount of computational resources involved in their training has increased significantly during recent years. Interactive training, ie an arrangement where the system continues to learn while interacting with the user,

[12] See, eg, A Clark, C Fox and S Lappin, *The Handbook of Computational Linguistics and Natural Language Processing* (Wiley, 2013).

[13] A Kelli, A Tavast, K Linden, R Birstonas, P Labropoulou, K Vider, I Kull, G Tavits, A Värv and V Mantrov, 'Impact of Legal Status of Data on Development of Data-Intensive Products: Example of Language Technologies', in *Legal Science: Functions, Significance and Future in Legal Systems II* (The University of Latvia Press, 2020) 383–400. doi.org/10.22364/iscflul.7.2.31.

utilising their utterances as additional training data, is also possible and is used in so-called recommender systems learning the user preferences and storing them in a user profile for giving personal recommendations.

See Figure 1 below for a simplified process diagram of creating a machine learning model.

#ArtworkB

Insert Figure 1.

#ArtworkE

Figure 1. Creating and using a machine learning model

Training data for language models consists of written text and speech recordings organised into corpora. Depending on the chosen machine learning technology and resources available to the model creator, these corpora may be manually or semi-automatically annotated or augmented with additional information layers. Creating the model starts from choosing, developing or customising the technology. As models grow, the training process itself has become so computationally intensive that reducing model sizes and training times is currently a major topic in machine learning research.[14]

Human intellectual contribution in the training process may include selecting and pre-processing the training data, hyper-parameter tuning, model testing and optimisation. It has been shown that hyper-parameter choices significantly impact the final results and even state of the art models may easily contain overlooked possibilities.[15] In case of unacceptable results, the process can be restarted at some earlier step. After reaching a model that is considered good enough, it can be deployed in a real product. A modern chatbot contains several such models performing various functions for understanding and generating language.

B. Modality of the interaction

Current chatbots communicate with the user in written or spoken form. For the spoken modality, speech recognition and speech synthesis models require human speech for training. Since a person's voice is personal data, model creation is subject to personal data protection regulations. Voice samples remain personal data unless they are transformed, compressed and

[14] See, eg X Jiao, Y Yin, L Shang, X Jiang, X Chen, L Li and FW Qun Liu, 'TinyBERT: Distilling BERT for Natural Language Understanding' (2020) arXiv:1909.10351 [cs.CL].

[15] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer and V Stoyanov, 'RoBERTa: A Robustly Optimized BERT Pretraining Approach' (2019) arXiv:1907.11692 [cs.CL].

stored as speaker-independent data samples, in which case the only remaining personal data may be in the content of the voice sample. In this case, the sample is similar to a text data sample.

An additional issue with voice-activated chatbots like Amazon Alexa or Google Home is the need to recognise the activation command. The chatbot must continually listen to speech around it to know when it is being addressed. Especially at home or elsewhere in the private sphere, such listening may cause justified privacy concerns. Of course, it is technically possible for the chatbot to ignore all other speech apart from its activation command, but knowing this may not be sufficient to alleviate the uneasiness associated with someone constantly listening.

C. Location of data processing

The technical detail amplifies privacy concerns that, especially speech models, are currently so resource-hungry that running them on edge devices such as phones, watches, smart speakers, etc, may not be realistic. It is easier to send the speech signal to the service provider's server to be processed using more computational power, returning the complete response to the user device. Guidelines on VVA also explains that since data is transferred to remote servers, the chatbot service provider as the data controller[16] needs to consider both the e-Privacy Directive[17] and General Data Protection Regulation. If future chatbots can provide services locally, the applicability of the e-Privacy Directive needs to be reassessed.[18]

These legal issues associated with such transmission and processing of data have been among the factors impeding the adoption of chatbots. This is part of the motivation to develop smaller and faster language models that can be run on less powerful user devices.

D. Length of memory

[16] Article 4(7) of the GDPR defines controller as 'the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data'.

[17] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [2002] OJ L201 as amended by Directive 2006/24/EC and Directive 2009/136/EC (e-Privacy Directive).

[18] Guidelines on VVA, 2, 12.

Considering that chatbots respond to user queries, they can be regarded as similar to search engines. Initially, the user query was the only piece of information that the search engine used for determining its response. The current default behaviour of mainstream search engines is different, however. Data such as the user's location and search history are used for improving the relevance of results.[19] Likewise, a chatbot may or may not store earlier utterances of the user with the aim of obtaining a more comprehensive view of the user and providing better reactions to future utterances.

Such storing itself must have some legal basis (eg consent, legitimate interest), considering that the system has no way of ensuring that user utterances do not contain personal data or copyrighted works. In this context, it is essential to follow the personal data processing principles such as transparency[20] (the user is informed about different processing activities), purpose limitation[21] (processed for a specific purpose), data minimisation (processed data is limited to what is strictly necessary) and storage limitation (data is kept as long as strictly needed) as provided by Article 5 of the General Data Protection regulation. Considering the storage limitation principle, Guidelines on VVA point out that some chatbot service providers keep personal data that requires deletion. This approach violates the principle of storage limitation. Personal data should not be retained for longer than it is necessary for the specific processing.[22]

Multiuser systems have the additional technical challenge of recognising which utterances originate from the same user. Four main approaches can currently be used to achieve this, each with their technical, usability-related and legal drawbacks: explicit registration and authentication of the user, browser session, cookies stored on the user's device, or IP address

[19] See, eg, S 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 (C-136/17 GC et al v CNIL)' (2019) 5/4 *European Data Protection Law Review* 561–567.

[20] See Art 29 Working Party. Guidelines on transparency under Regulation 2016/679. Adopted on 29 November 2017. As last Revised and Adopted on 11 April 2018. ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.

[21] See Art 29 Working Party. Opinion 03/2013 on purpose limitation. Adopted on 2 April 2013. ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[22] Guidelines on VVA, 3.

of the device. The inclusion of the latter among personal data[23] has reduced but not completely obliterated its use.

Guidelines on VVA require chatbot designers to consider the use of technologies for filtering out unnecessary background noise (eg a third person's speech) to ensure that only the user's voice is recorded.[24]

E. Purpose of use of data

According to the Guidelines on VVA, the four most common purposes for which chatbots process personal data are: 1) executing requests; 2) improving the machine learning model; 3) biometric identification; and 4) profiling for personalised content or advertising. Of these four, if

> data is processed in order to execute the user's requests, i.e. as strictly necessary in order to provide a service requested by the user, data controllers are exempted from the requirement of prior consent under Article 5(3) the e-Privacy Directive.[25] Conversely, such consent as required by Article 5(3) of the e-Privacy Directive would be necessary for the storing or gaining of access to information for any purpose other than executing users' request.[26]

III. Legal framework for building chatbots

Chatbots do not contain different or separate components from NLP or entail other legal issues. The authors have previously discussed legal problems in NLP.[27] In this section, the authors

---

[23] According to the EU case law (Case C-70/10 *Scarlet Extended SA v SABAM* [2011] ECR I-11959) IP addresses constitute personal data since they allow to identify the users. It was later specified that 'a dynamic IP address registered by an online media services provider when a person accesses a website that the provider makes accessible to the public constitutes personal data' (Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* [2016] Digital reports.

[24] Guidelines on VVA, 3.

[25] Article 5(3) e-Privacy Directive read as follows: 'Member States shall ensure that the storing of information, or the gaining of access to information already stored, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned has given his or her consent, having been provided with clear and comprehensive information, in accordance with Directive 95/46/EC, inter alia, about the purposes of the processing. This shall not prevent any technical storage or access for the sole purpose of carrying out the transmission of a communication over an electronic communications network, or as strictly necessary in order for the provider of an information society service explicitly requested by the subscriber or user to provide the service'.

[26] Guidelines on VVA, 2-3.

[27] A Kelli, A Tavast, K Linden, R Birstonas, P Labropoulou, K Vider, I Kull, G Tavits, A Värv and V Mantrov, 'Impact of Legal Status of Data on Development of Data-Intensive Products: Example of Language Technologies', in *Legal Science: Functions, Significance and Future in Legal Systems II* (The University of Latvia

address challenges relating to legal restrictions on language data used for building chatbots, lawful bases to use the data and the legal status of the models.

A. Data used to build a chatbot

The creation of chatbots requires the use of training data (language data). Without doubt, it is preferable and more convenient to use data that does not have any legal restrictions. Anonymous data sis preferable since it is outside the scope of the GDPR.[28] It is also possible to anonymise personal data.[29]

However, it is often unavoidable to use data containing personal data or copyrighted content. Article 4(1) of the GDPR defines personal data[30] as follows: 'any information relating to an identified or identifiable natural person ('data subject')'. It is a real challenge to draw a line between data relating to an indirectly identifiable natural person and anonymous data.[31]

The GDPR also regulates special categories of personal data, for which processing is even more restricted.[32] Guidelines on VVA acknowledges that data processed by chatbots could be sensitive since 'It may carry personal data both in its content (meaning of the spoken text) and its meta-information (sex or age of the speaker etc.).'[33]

As a general rule, the GDPR protects the personal data of a living person. Recital 27 of the GDPR explains that the regulation does not apply to the personal data of deceased persons.

Press, 2020) 383–400. doi.org/10.22364/iscflul.7.2.31; A Tavast, H Pisuke and A Kelli, 'Õiguslikud väljakutsed ja võimalikud lahendused keeleressursside arendamisel (Legal Challenges and Possible Solutions in Developing Language Resources)' (2013) 9 *Eesti Rakenduslingvistika Ühingu aastaraamat* 317–332. dx.doi.org/10.5128/ERYa9.20.

[28] Recital 26 of the GDPR explains: 'The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable'.

[29] See WP29. Opinion 05/2014 on Anonymisation Techniques Adopted on 10 April 2014. ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[30] For further explanation of the concept of personal data, see Art 29 Working Party (WP29). Opinion 4/2007 on the concept of personal data. Adopted on 20th June. ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

[31] See G Spindler and P Schmechel, 'Personal Data and Encryption in the European General Data Protection Regulation' (2016) 7 *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*.

[32] The GDPR Art 9(1) defines special categories of personal data as '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'.

[33] Guidelines on VVA, 12.

However, EU Member States may regulate this issue. For instance, according to the Estonian Personal Data Protection Act 2018, section 9(2), personal data is protected for ten years after the death of the data subjects (20 years in case of minors).

The definition of personal data is rather extensive, and it covers a variety of data potentially relevant for chatbots, such as a person's name,[34] direct and indirect identifiers and a person's voice. The human voice, of which samples are needed to train models for chatbots, is a particularly interesting legal phenomenon.[35] Article 4(14) of the GDPR defines it as biometric data,[36] which in Article 9(1) is listed as special categories of personal data in case it is processed 'for the purpose of uniquely identifying a natural person'. The question is whether the human voice as such is covered by the special categories of personal data. The situation is analogous to photographs of people as a person's image also contains biometric data (physical characteristics) named as special categories of personal data (GDPR Article 9 (1)). Recital 51 of the GDPR explains that:

> The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.

The authors are of the opinion that the same approach applies to voice as well. Unless the voice is used to identify an individual, it is not in the special categories of personal data, and its processing is not subject to the stricter requirements applicable to special categories of personal data.[37]

Although it is recommendable to use non-copyrighted content, the reality is that data used to build chatbots is often copyright protected. Drawing a line between copyrightable and non-copyrightable content is similarly challenging, as is the case with personal and non-

---

[34] A person's name is considered personal data already by early EU case law. See Case C-101/01 Criminal proceedings against Bodil Lindqvist [2003] ECR I-12971.

[35] See I Ilin and A Kelli, 'The Use of Human Voice and Speech for the Development of Language Technologies: the EU and Russian Data Protection Law Perspectives' (2020) 29 *Juridica International* 71−85. www.juridicainternational.eu/article_full.php?uri=2020_29_the_use_of_human_voice_and_speech_for_development_of_language_technologies_the_eu_and_russia.

[36] According to the GDPR Art 4 (14) ''biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data'.

[37] Guidelines on VVA also support the position that unless voice is used to identify an individual, it is not considered as special categories of personal data.

personal data. Copyright law protects works. According to Article 2(1) of the Berne Convention, 'The expression "literary and artistic works" shall include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression.'[38] The concept of copyright protection is addressed in the EU case law as well. The European Court of Justice (ECJ) has held that copyright applies 'only in relation to a subject-matter which is original in the sense that it is its author's own intellectual creation'.[39] According to the ECJ the originality exists 'if the author was able to express his creative abilities in the production of the work by making free and creative choices'.[40] Intellectual creations cannot be classified as works when there is no room for creative freedom for the purposes of copyright.[41]

When it comes to the development of chatbots, our emphasis is on works that consist of words. One word is not copyright protected since there is no creativity. The intriguing question is how many words it takes to qualify as a copyright-protected work.[42] The ECJ has also analysed the issue and put forward the following explanation:

> words which, considered in isolation, are not as such an intellectual creation of the author who employs them. It is only through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation.

At the same time, the court suggested that 11 consecutive words are protected by copyright if that extract contains an element of the work, which, as such, expresses the author's own intellectual creation.[43]

---

[38] European countries have a similar approach. For instance, s 4(2) of the Estonian Copyright Act 1992 defines works as 'any original results in the literary, artistic or scientific domain which are expressed in an objective form and can be perceived and reproduced in this form either directly or by means of technical devices. A work is original if it is the author's own intellectual creation'.

[39] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-06569, para 37.

[40] Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH and Others* [2011] ECR I-12533, para 89.

[41] Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd and Others v QC Leisure and Others* [2011] ECR I-09083, para 98.

[42] For further discussion, see P Kamocki, 'When Size Matters. Legal Perspective(s) on N-grams'. Proceedings of CLARIN Annual Conference 2020. 05–07 October 2020. Virtual Edition. C Navarretta and M Eskevich (eds) (CLARIN 2020) 166–169. office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.

[43] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-06569, paras 45, 48.

In case speech is used, there is a need to consider related rights[44] to copyright as well.[45] In this situation, it is important to distinguish between the author of the work and the performer of the work. When works are protected by copyright, then the performances are protected by related rights.[46] A speech can be considered a performance. If the speech is pre-recorded, the phonogram producer's rights have to be respected as well.[47]

Copyrighted content is sometimes extracted from databases. We can distinguish copyright protected and sui generis databases. According to Article 7 of the Database Directive,[48] the maker of a sui generis database acquires rights to allow and forbid the extraction and re-utilisation of the database or its substantial part if the maker

> shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilisation of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database.

Sui generis database rights arise from the investment in the development of a database, not in the creation of the data in it. According to the EU case law, investment refers to 'the resources used to seek out existing independent materials and collect them in the database. It does not cover the resources used for the creation of materials which make up the contents of a database'.[49]

It should be mentioned that the DSM Directive might introduce a potentially relevant right for language research affecting the creation of language models used for chatbots.

---

[44] Sometimes related rights are referred to as neighbouring rights. The nature of the related rights is that they contribute to dissemination of copyright protected works (performer's rights, phonogram producer's rights) or protect investment (rights of maker of sui generis databases).

[45] In legal practice, both type of rights are usually transferred to third parties (more often to legal entities) who are called rightholders.

[46] Article 2(a) of the International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations (Rome 26 October 1961) defines performers as 'actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, or otherwise perform literary or artistic works'.

[47] Article 2(c) of the International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations defines the producer of phonograms as 'the person who, or the legal entity which, first fixes the sounds of a performance or other sounds'.

[48] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive) OJ L77.

[49] C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd* [2004] ECR I-10415. For further analysis, see A Kelli, A Tavast, K Lindén, K Vider, R Birštonas, P Labropoulou, I Kull, G Tavits, A Värv, P Stranák and J Hajic, 'The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies', in K Simov and M Eskevich (eds), *Selected Papers from the CLARIN Annual Conference 2019* (Linköping University Electronic Press, 2020) 53−65. ep.liu.se/en/conference-article.aspx?series=ecp&issue=172&Article_No=8.

According to Article 15 of the DSM Directive, publishers of press publications shall have the reproduction right and the making available right for the online use of their press publications by information society service providers. The impact of this right remains to be seen.

The acknowledgement that training data required to build a chatbot often contains personal data and content protected by copyright and related rights does not exclude its use. The use is possible, but it has to follow the requirements set forth by copyright and data protection laws analysed in the following section.

B. Legal basis for building a chatbot

There are two different scenarios for building a chatbot. This distinction relies on the chatbot's life cycle. According to the first scenario (the initial creation), data containing personal data and copyrighted content is used from different sources to create a language model for a chatbot.[50] The second scenario relates to the improvement of the model during the use of a chatbot. The Guidelines on VVA explain the need for improvement as follows

> [t]he accents and variations of human speech are vast. While all VVAs are functional once out of the box, their performance can improve by adjusting them to the specific characteristics of users' speech.

To improve machine learning methods, chatbot designers wish to have access to and process data (eg, voice snippets) relating to the use of the device. Article 5(3) of the e-Privacy Directive requires consent for 'gaining of access to information for any purpose other than executing users' request'.[51] The authors concentrate on the initial creation since the improvement of the existing chatbot relies on standard terms of use and the user's consent to process his personal data.

From a legal point of view, the creation of models involves text and data mining TDM. Article 2(2) of the DSM Directive defines text and data mining as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'.

---

[50] Theoretically we could distinguish two cases here: 1) language model is developed for a specific chatbot; 2) language model is acquired from a third party. However, in the second case, the creation of the acquired language model had to be based on some legal bases which makes the case identical to the first case. Therefore, the second case is not discussed.

[51] Guidelines on VVA, 22, 10, 12.

Text and data mining has a different meaning in personal data protection and copyright context. From the personal data protection perspective, text and data mining constitutes such processing of personal data that needs a legal basis. The reason is that the GDPR defines processing so extensively that any operation performed on personal data is processing.[52]

From the copyright perspective, text and data mining as such is not copyright relevant.[53] It is comparable to reading a book. However, a legal basis is needed to copy the data containing copyrighted content for subsequent text and data mining.

Generally speaking, the use of personal data and copyrighted content can be based on permission or some other legal bases.

According to Article 6(1) of the GDPR, consent is a legal basis for processing personal data. Article 4(11) of the GDPR defines the data subject consent as

> any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.

Article 7 of the GDPR describes conditions for consent such as the burden of proof, language and distinguishability from the other matters, withdrawal of consent and criteria for the assessment of consent.[54] Although consent potentially guarantees a high level of protection of the rights and freedoms of the data subject, its acquisition is not always possible. For instance, the creation of language models could require the use of a high number of blog posts or video recordings or, many years ago collected language data (legacy data). Big data[55] poses similar problems.[56] In the referred cases, it is not realistic to acquire the data subject's consent

---

[52] Article 4(2) of the 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'.

[53] This is also emphasised in Recital 9 of the DSM Directive which states 'Text and data mining can also be carried out in relation to mere facts or data that are not protected by copyright, and in such instances no authorisation is required under copyright law'.

[54] For explanation of the concept of consent, see European Data Protection Board. Guidelines 05/2020 on consent under Regulation 2016/679. Version 1.1. Adopted on 4 May 2020. edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf.

[55] For an explanation of the concept of big data, see R Kitchin and G McArdle, 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets' (2016) *Big Data & Society* 1–10. journals.sagepub.com/doi/full/10.1177/2053951716631130

[56] It is pointed out in legal literature that big data and personal data protection controversies reveal 'difficulties in finding a legitimate ground for processing or acquiring consent from the data subject' - M Oostveen,

due to an extremely high administrative burden or lack of contact data. Therefore, some other legal ground is needed.

The GDPR provides two other legal grounds besides consent which could potentially be relied upon.[57] The first possible legal ground is the performance of a task carried out in the public interest.[58] The reliance on this legal ground is based on the logic that research organisations and universities conduct research in the public interest. This means that language data has to be processed to develop language models by the university. Private entities cannot use this legal ground. However, Recital 159 of the GDPR explains that

> the processing of personal data for scientific research purposes should be interpreted in a broad manner including, for example, technological development and demonstration, fundamental research, applied research and privately funded research.

This means that a public-private partnership is possible and the private sector can develop language models in collaboration with research organisations.

Another legal ground for processing personal data is legitimate interests.[59] There are no limitations as to potential controllers as is the case with research organisations acting in public interest. Every entity and individual can process with a legitimate interest. However, there is no clarity as to its exact scope, so the legitimate interest needs to be motivated for each case.[60]

The acquisition of permission from the rightholder to use copyrighted content faces similar challenges as described above concerning getting consent to process personal data. Therefore, there is a need to rely on some other legal basis (ie copyright exceptions) to copy[61] copyrighted content for TDM.

'Identifiability and the applicability of data protection to big data' (2016) 6 (4) *International Data Privacy Law* 309.

[57] For further discussion, see A Kelli, K Lindén, K Vider, P Kamocki, R Birštonas, S Calamai, P Labropoulou, M Gavriilidou and P Stranák, 'Processing personal data without the consent of the data subject for the development and use of language resources', in I Skadina and M Eskevich (eds), *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018* (Linköping University Electronic Press, 2019) 72–82. ep.liu.se/ecp/article.asp?issue=159&article=008&volume=.

[58] The GDPR Art 6(1)(e).

[59] ibid Art 6(1)(f).

[60] For a further explanation of the concept of legitimate interest, see WP29. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Adopted on 9 April 2014. ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

[61] It should be mentioned that TDM mainly concerns the reproduction right. Article 2 of the InfoSoc Directive defines reproduction right as 'exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part'.

Before adopting the DSM Directive, the InfoSoc Directive constituted the copyright framework for making copies of copyrighted content for text and data mining.[62] The potential exceptions are the private use exception, the quotation right, the temporary reproduction right and the research exception.[63]

The private use exception is meant for natural persons for private use[64] and has limited impact on building a chatbot. Anyone can rely on the quotation right and the temporary reproduction right.[65] Theoretically, the quotation right could be a suitable legal basis to use copyrighted content since it does not exclude commercial purposes. However, the scope of the quotation right is somewhat limited. According to Article 5(3)(d), quotations are allowed

> for purposes such as criticism or review, provided that they relate to a work or other subject-matter which has already been lawfully made available to the public, that, unless this turns out to be impossible, the source, including the author's name, is indicated, and that their use is in accordance with fair practice, and to the extent required by the specific purpose.

The narrow scope has been reinforced by the EU case law as well. The European Court of Justice has explained that 'the user of a protected work wishing to rely on the quotation exception must therefore have the intention of entering into "dialogue" with that work'.[66] Since content protected by copyright and related rights is used as raw material to develop language models and there is no 'dialogue' as required by the EU acquis, then the quotation right is not applicable.

The temporary reproduction right[67] might be a legal basis to copy copyrighted content for TDM to create language models. Its scope covers temporary acts of reproductions which

---

[62] For the discussion on the general framework of TDM, see JP Triaille, J de Meeûs d'Argenteuil and A de Francquen, 'Study on the legal framework of text and data mining (TDM)' (March 2014). op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en.

[63] See A Kelli, A Tavast, K Linden, R Birstonas, P Labropoulou, K Vider, I Kull, G Tavits, A Värv and V Mantrov, 'Impact of Legal Status of Data on Development of Data-Intensive Products: Example of Language Technologies', *Legal Science: Functions, Significance and Future in Legal Systems II* (The University of Latvia Press, 2020) 383–400. doi.org/10.22364/iscflul.7.2.31.

[64] InfoSoc Directive Art 5(2)(b).

[65] ibid Art 5(3)(d) and 5(1).

[66] Case C-476/17 *Pelham GmbH, Moses Pelham, Martin Haas v Ralf Hütter, Florian Schneider-Esleben* [2019] Digital reports, para 71.

[67] InfoSoc Directive Art 5(1) reads as follows: 'Temporary acts of reproduction referred to in Article 2, which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable:

(a) a transmission in a network between third parties by an intermediary, or

are an integral part of a technological process.[68] Recital 9 of the DSM Directive also suggests that the temporary reproduction right 'should continue to apply to text and data mining techniques that do not involve the making of copies beyond the scope of that exception'. However, some experts warn that 'a considerable level of uncertainty surrounds the applicability of the exception for temporary uses of Art. 5(1) InfoSoc and a proper analysis of each case should be performed before relying on it'.[69]

According to Article 5(3) clause a) of the InfoSoc Directive, the EU Member States may provide for exceptions to the reproduction right for 'scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved'. Although the research exception excludes commercial purposes, it can be used to copy content for TDM. Legal commentators also emphasise that 'the exception is useful and relevant for TDM'.[70]

The DSM Directive creates a specific framework for text and data mining. The regulation covers the TDM exception for research purposes which is meant for research organisations and cultural heritage institutions, and the general TDM exception meant for everyone.[71] Both exceptions limit the reproduction right and the right to make extractions from a database. The reliance on the general TDM exception can be excluded by the rightholder.[72] At the same time, any contractual provision limiting the TDM for research purposes is unenforceable.[73] According to Article 3(2) of the DSM Directive, the TDM exception for research purposes also allows retaining a copy of protected content 'for the purposes of

---

(b) a lawful use

of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2'.

[68] The temporary reproduction right is also addressed in the EU case law focusing on similar technological process, see Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* [2009] ECR I-06569; Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* [2012] Digital reports.

[69] RE de Castilho, G Dore, T Margoni, P Labropoulou and I Gurevych, 'A Legal Perspective on Training Models for Natural Language Processing', in N Calzolari et al (eds) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018) 1273–1274. www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf.

[70] JP Triaille, J de Meeûs d'Argenteuil and A de Francquen, 'Study on the legal framework of text and data mining (TDM)'. (March 2014), 61. op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en.

[71] The DSM Directive Arts 3 and 4.

[72] ibid Art 4(3).

[73] The DSM Directive Art 7(1).

scientific research, including for the verification of research results'. Article 4(2) of the DSM Directive regulating the general TDM exception allows keeping the copy of the content as 'for as long as is necessary for the purposes of text and data mining'. Therefore, it can be concluded that the TDM exception for research purposes is more favourable than the general TDM exception.

The TDM exception for research purposes does not require that TDM to create language models must be done in isolation by research organisations. Recital 11 of the DSM Directive encourages research organisations to collaborate with the private sector and carry out TDM in public-private partnerships.

The analysis revealed that from the legal point of view, there are fewer legal restrictions on the use of content containing copyrighted works and personal data when it is done by research organisations. At the same time, the legal framework favours public-private partnership.

C. Legal status of language models

The authors have previously analysed whether language models are subject to the same legal restrictions as the data used to create them and concluded that it is usually not the case.[74]

From the copyright perspective, the main issue is whether the language model contains copyrighted content. As a general rule, models do not include copyrighted works used for their creation since they rely on small snippets which are not original enough to be copyright-protected. It is also found in legal literature that a model 'does not reproduce the original (corpora) nor reveals 'its individuality'.[75] If models contain copyrighted content, copyright requirements need to be followed when using the model. In practical terms, it means that the rightholder's permission has to be acquired.

---

[74] A Kelli, A Tavast, K Lindén, K Vider, R Birštonas, P Labropoulou, I Kull, G Tavits, A Värv, P Stranák and J Hajic, 'The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies', in K Simov and M Eskevich (eds), *Selected Papers from the CLARIN Annual Conference 2019* (Linköping University Electronic, Press 2020) 53−65. ep.liu.se/en/conference-article.aspx?series=ecp&issue=172&Article_No=8.

[75] RE de Castilho, G Dore, T Margoni, P Labropoulou and I Gurevych, 'A Legal Perspective on Training Models for Natural Language Processing', in N Calzolari et al. (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018) 1273–1274. www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf.

The same question needs to be answered regarding personal data. In case there is no personal data, the person who created the model is free to use it at his discretion (share for free, sell, etc). It should be borne in mind that the personal data rules do not apply to anonymous data.[76] The success of anonymisation depends on the type of data.[77] The Guidelines on VVA referring to scientific literature[78] warn that there are risks of re-identifying persons in some machine learning models. Therefore, mitigation measures need to be applied to reduce the re-identification risk to an acceptable threshold.[79]

If a model contains personal data, there must be a legal basis to use the model in a chatbot. Depending on the context, the data subject's consent or the legitimate interest of the user could be the legal base.

There could be an interesting legal case when a language model was created using language data without a legal basis. The injured party has the right to compensation, and administrative fines can be imposed on the violator in case of a GDPR violation.[80] The Enforcement Directive,[81] which establishes the general framework for intellectual property enforcement, entitles the injured party to claim damages as well. The injured party can request the termination of illegal processing of personal data and unlawful use of copyrighted works. However, the question is what happens to a model which was created illegally, but the model itself does not contain any personal data or copyrighted content. There was an analysis in OneZero of a US case when a company used private photos to train facial recognition algorithms. The Federal Trade Commission required the company to delete all photos and algorithms which were developed using the data.[82] It remains to be seen if the EU regulatory framework will be interpreted in a way allowing to require the deletion of illegally built models.

---

[76] Recital 26 of the GDPR.

[77] It is suggested that 'Anonymizing voice recordings is specially challenging, as it is possible to identify users through the content of the message itself and the characteristics of voice itself' – The Guidelines on VVA, 26.

[78] M Veale, R Binns and L Edwards, 'Algorithms that remember: model inversion attacks and data protection law' (2018) *Philosophical Transactions of Royal Society A*. doi.org/10.1098/rsta.2018.0083; N Carlini et al, 'Extracting Training Data from Large Language Models' (2020). www.researchgate.net/publication/347125123_Extracting_Training_Data_from_Large_Language_Models.

[79] The Guidelines on VVA, 25–26.

[80] The GDPR Art 82 and 83.

[81] Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (Enforcement Directive) [2004] OJ L157.

[82] D Gershgorn, 'The FTC Forced a Misbehaving A.I. Company to Delete Its Algorithm' (2021) *OneZero*. onezero.medium.com/the-ftc-forced-a-misbehaving-a-i-company-to-delete-its-algorithm-124d9f7e0307.

Theoretically, the measure to demand the deletion of a model could be applicable in cases when due to lack of legal competencies or intentionally, a model is created without a proper legal basis. Therefore, it becomes even more crucial to follow copyright and personal data protection rules when building models.

IV. Conclusion

Chatbots rely on language models. Models may be trained on language data containing copyrighted material and personal data. From a legal perspective, training involves text and data mining (TDM). TDM itself is not copyright relevant. This means that performing text and data mining is not regulated by copyright, and it does not require any legal basis. However, to copy copyrighted content for TDM, there has to be a legal basis. The potential legal base could be the rightholder's permission or a copyright exception. The InfoSoc Directive provides the temporary reproduction right and the research exception as possible legal bases. The DSM Directive introduces two frameworks for text and data mining: 1) TDM exception for research purposes; and 2) a general TDM exception.

The GDPR defines the processing of personal data so extensively that it covers all possible operations on the data. Therefore, there has to be a legal basis for the collection and analysis (TDM) of personal data. The potential legal case could be the data subject's consent, research in the public interest or legitimate interest.

Considering copyright and personal data protection in combination, it becomes apparent that the regulatory framework for the creation of language models is more favourable for research organisations. Both legal frameworks favour public-private collaboration, which means that private companies can cooperate with research organisations to build language models within the more favourable framework meant for research.

When it comes to models, then the model training can typically be performed so that it is impossible to re-create the training data from the model, and the model does not contain original portions of works included in the training data. After meeting these conditions, the model is a new independent work, disconnected from the training material in terms of copyright. Models also contain no personal data if care is taken that potential identifiers such as names, social security numbers, addresses etc, are stored only in combinations that cannot

identify a real person. In case personal data remains in the models, there has to be a legal basis for its processing.

An intriguing issue relates to a case where a model is trained on language data containing copyrighted content and personal data without a proper legal basis. Even if the built model does not have any personal data or copyrighted content, there is a risk that the model has to be deleted.