**ARTICLE**

# Potentials and Challenges of Large Language Models (LLMs) in the Context of Administrative Decision-Making

Paulina Jo Pesch🔟

School of Law, FAU Erlangen-Nürnberg, Erlangen, Germany
Email: paulina.pesch@fau.de

## Abstract

Large Language Models (LLMs) could facilitate both more efficient administrative decision-making on the one hand, and better access to legal explanations and remedies to individuals concerned by administrative decisions on the other hand. However, it is an open research question of how performant such domain-specific models could be. Furthermore, they pose legal challenges, touching especially upon administrative law, fundamental rights, data protection law, AI regulation, and copyright law. The article provides an introduction into LLMs, outlines potential use cases for such models in the context of administrative decisions, and presents a non-exhaustive introduction to practical and legal challenges that require in-depth interdisciplinary research. A focus lies on open practical and legal challenges with respect to legal reasoning through LLMs. The article points out under which circumstances administrations can fulfil their duty to provide reasons with LLM-generated reasons. It highlights the importance of human oversight and the need to design LLM-based systems in a way that enables users such as administrative decision-makers to effectively oversee them. Furthermore, the article addresses the protection of training data and trade-offs with model performance, bias prevention and explainability to highlight the need for interdisciplinary research projects.

**Keywords:** administrative decision-making; artificial intelligence (AI); automated decision-making; large language models (LLMs)

## I. Introduction

With its release in November 2022, ChatGPT[1] has caused a hype around Large Language Models (LLMs) and fuelled the discussion around the chances and risks of AI models.[2] Not only private end users and enterprises but also authorities are exploring general-purpose models such as ChatGPT, Claude[3] or Gemini.[4] They are experimenting with their own domain-specific LLMs

---

[1] Open AI, ChatGPT <https://chat.openai.com/> accessed 10 December 2024.
[2] Yogesh K. Dwivedi et al., "Opinion Paper: 'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy" (2023) 71 International Journal of Management 102642; Krzysztof Wach et al., "The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT" (2023) 11/2 Entreprenurial Business and Economics Review, 7; Laura Weidinger et al., "Taxonomy of Risks Posed by Language Models" (2022) 2022 Proceedings FaccT, 214.
[3] Anthropic Claude, <https://claude.ai> accessed 10 December 2024.
[4] Google Gemini, <https://gemini.google.com/> accessed 10 December 2024 (formerly known as "Bard").

created by means of fine-tuning pre-trained models with domain-specific training data.[5,6] Also, there are initiatives to use such models to help individuals to understand decisions concerning them and effectively take legal action.[7] While the models could increase the efficiency of administrative decision-making procedures and make legal aid available to vulnerable groups, there remain concerns about LLMs' functional limitations and compliance with legal requirements especially in the fields of data protection[8] and administrative law. Existing sources on the practical and legal issues are limited insofar as technical studies on existing LLMs[9] do not generalise across different models and domains, and legal literature on automated decision-making largely refers to automated systems or machine learning models in general[10] rather than to LLMs in specific.

This article provides an overview of the potential and challenges of LLMs in the context of administrative decision-making. As the analysis requires a sufficient understanding of the technical functioning of LLMs, section II provides a brief introduction into the functioning of existing LLMs. Section III outlines potential use cases for LLMs in the context of administrative decision-making on both sides, the authorities on the one hand and concerned individuals on the other hand. Selected practical and legal challenges are discussed in section IV. Section V draws a conclusion and an outlook to future research.

## II. Existing large language models (LLMs) in a Nutshell

Understanding LLMs requires an idea of the underlying architecture and the training (section II.1). Also, it is necessary to delineate non-generative from generative models (section II.2). Furthermore, the use of personal data for the training of the models and the storage of training data in the model parameters in such a way that the training data can be extracted ("memorisation") must be taken into account (section II.3). The introduction is based on existing general-purpose models, ie, models that have the capability of serving

---

[5] On the term fine-tuning and for example in the legal domain Davide Liga and Livio Robaldo, "Fine-Tuning GPT-3 for Legal Rule Classification' 2023 (51) CSLR 105864.

[6] See for example Beck Aktuell, "NRW und Bayern entwickeln 'ChatGPT-Analogon' für die Justiz" (2023) <https://rsw.beck.de/aktuell/daily/meldung/detail/nrw-und-bayern-entwickeln-chatgpt-analogon-fuer-die-justiz> accessed 10 December 2024: North Rhine-Westphalia and Bavaria are developing a generative LLM for courts; Osmond Chia, "Public Officers Can Use ChatGPT and Similar AI, But Must Take Responsibility for Their Work: MCI" (2023) <https://www.straitstimes.com/tech/public-officers-allowed-to-use-chatgpt-and-other-ai-but-must-take-responsibility-for-work-mci> Luke Taylor, "Colombian Judge Says He Used ChatGPT in Ruling' (2023) <https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling> accessed 10 December 2024; Fedscoop, 'Government Gears Up to Embrace Generative AI' (2023) <https://fedscoop.com/government-gears-up-to-embrace-generative-ai/> accessed 10 December 2024.

[7] See, for example, Hack To The Rescue, Challenges (2023) via Wayback Machine, <https://web.archive.org/web/20231202225502/https://hacktotherescue.org/generativeai/challenges> accessed 10 December 2024, "Auto-translate the legal documents for refugees into simple language" and "Automated Human Rights appeal creator to empower Refugees and Asylum seekers to fight for their rights".

[8] On the classification of LLMs as personal data under the GDPR and data accuracy requirements for the models and their outputs Paulina Pesch and Rainer Böhme, "Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen – ChatGPT & Co. unter der DSGVO" 2023 (12) MMR 917. On LLM-driven applicant selection under Art 22 GDPR, Art 14 AI Act Johannes Kätscher and Paulina Jo Pesch, "Automatisierte Entscheidungsfindung mittels großer Sprachmodelle (LLM) im Beschäftigtenkontext – Interview mit einem Chatbot" 2024 KIR 46.

[9] See in particular the studies on training data extraction, section II.3, and the attempts to use LLMs for legal reasoning, section IV.1.b with n 64 f.

[10] See, for example, Melanie Fink and Michèle Finck, "Reasoned A(I)dministration: Explanation Requirements in EU Law and the Automation of Public Administration" (2022) 3 ELR 376 [380] with further references.

a variety of purposes[11] such as ChatGPT, Claude, Gemini or BERT which only allow cautious conclusions about the potential properties of future models.

### 1. Architecture and training

LLMs are based on machine learning, the – at present – most relevant subfield of artificial intelligence (AI). The architecture of the models is referred to as artificial neural networks (ANNs), complex structures of mathematical instructions represented in a high number of nodes and connections between them.[12] An untrained ANN has unspecified (respective random) parameters that serve as placeholders for information. A model with specified parameters is built through training with usually large amounts of training data.

Training an LLM requires training data in the form of text. General-purpose models such as ChatGPT, Claude, Gemini or BERT, for example, are trained not only but also with large amounts of text data that is publicly available online, eg, books, research articles, Wikipedia articles and other websites.[13] For the training, HTML code and URLs[14] are usually eliminated from the texts,[15] and the cleaned texts are broken down to "tokens", ie, sentences, parts of sentences, words, parts of words or even single characters. During the first phase of the training, the model is iteratively fed with incomplete sentences and predicts missing word(s), eg, the next one.[16] When this phase of training is completed, a model can generate outputs with correct syntax. To perform sufficiently with respect to semantics, the model is trained on questions through reinforcement, either based on question-answer-pair databases,[17] or on human feedback.[18] However, generative LLMs that generate syntactically and semantically correct outputs, do not understand grammar rules or the meaning of words.[19] Instead, existing models merely operate with statistical relationships between words. As machine learning models, in contrast to rule-based AI

[11] Cf Art 3(66) Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) P9_TA(2024)0138.

[12] Anders Krogh, "What Are Artificial Neural Networks?" (2008) 26/2 Nature Biotechnology 195.

[13] For ChatGPT cf Open AI FAQ, "How ChatGPT and Our Language Models Are Developed" <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed> accessed 10 December 2024. On BERT Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" v2 (2019) 5 arXiv:1810.04805. For LaMDA and Bard that have been trained on Google's Infinset dataset see Joe Jacon, "What Sites Were Used for Training Google Bard AI?" <https://medium.com/@taureanjoe/what-sites-were-used-for-training-google-bard-ai-1216600f452d> accessed 10 December 2024; Romal Thoppilan et al., "LaMDA: Language Models for Dialog Applications" (2022) 47arXiv:2201.08239v3.

[14] For LLMs connected to search engines, however, it is possible that the model store certain URLs to access up-to-date information to respond to prompts. Alternatively, LLMs can derive search strings from text prompts and by this access up-to-date information.

[15] Nicholas Carlini et al., "Extracting Training Data from Large Language Models" (2021) 30th USENIX Security Symposium 2635.

[16] On BERT Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" v2 (2019) 4 arXiv:1810.04805; on ChatGPT Open AI, "Aligning Language Models to Follow Instructions" <https://openai.com/research/instruction-following> accessed 10 December 2024.

[17] Such databases are rare but have been used to train BERT, Skada Vivek, "Fine-Tune Transformer Models For Question Answering On Custom Data" (2022) <https://towardsdatascience.com/fine-tune-transformer-models-for-question-answering-on-custom-data-513eaac37a80> accessed 10 December 2024.

[18] On human feedback provided to ChatGPT Tianyu Wu et al., "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development" (2023) 10/5 IEEE/CAA Journal of Automatica Sinica 1122 [1126 f.].

[19] On the lack of understanding also see John Morison and Tomas McInerney, "When Should a Computer Decide? Judicial Decision-Making in the Age of Automation, Algorithms and Generative Artificial Intelligence" in S Turenne and M Moussa (eds), *Research Handbook on Judging and the Judiciary* (Cheltenham, Edward Elgar Publishing Ltd 2024) p 27.

systems,[20] they derive patterns from the training data without explicit, interpretable rules. LLMs "learn", based on the training data, which words in which order are likely in a specific context and, thereby, adjust their parameters such that it can generate highly probable word combinations.[21] Because LLMs are opaque machine-learning models, it is not possible to determine how a particular input led to a particular output, ie, to trace elements from their output to elements from the input with certainty.

## 2. Generative vs. non-generative LLMs

Non-generative LLMs such as BERT are used to summarise, translate or classify text. Generative models such as ChatGPT, Claude, or Gemini can perform all these tasks. In addition, they are able to creatively generate novel texts. The goal of creating novel texts, ie, texts that to some extent differ from the statistical patterns in the training data, seems to contradict the training goal of the models to generate highly likely outputs as these are per se similar to the training data. To create something new, generative LLMs use so-called temperature parameters[22]: Setting the temperature parameters low results in more deterministic outputs, ie, more likely word combinations that more closely follow the patterns and probabilities observed in the training data. Higher temperatures increase the variance of the generated outputs, leading to more diverse and creative responses that deviate further from the training data. This is because temperature parameters determine to which extent a generative model mixes statistically less likely words in its outputs. However, adjusting temperature parameters is tricky: Even a slightly too high temperature causes a model to generate meaningless gibberish. However, even "well-tempered" generative LLMs that were trained on correct data are not fully reliable as they often fabricate incorrect information ("hallucinations").[23]

As this article explores the use of LLMs to generate complete administrative decisions, analyses or complaints (see section III), it focuses on generative models. For the following sections the term "LLMs" without additional specification refers to generative LLMs.

## 3. The use and "Memorisation" of personal data in the training of LLMs

Training data for existing LLMs contain personal data, for example, the names of authors of research articles and books, and of any real persons referred to in books, research articles, Wikipedia articles and texts from other websites. It is usually impracticable if not impossible to completely clean the large amounts of training data from personal data. Also, the models are specifically trained with personal data such as names and addresses so they "learn" how this information fits in texts.[24] Future breakthroughs in anonymisation

---

[20] On rule-based systems see Peter Parycek, Verena Schmid and Anna-Sophie Novak, "Artificial Intelligence (AI) and Automation in Administrative Provedures: Potentials, Limitations, and Framework Conditions" 2024 (15) Journal of the Knowledge Economy 8390 [8391 f, 8400 ff]. It is also possible to combine rule-based and machine learning approaches, see, for example, Jan Leismann, Chao Wang and Sven Mayer, "Comparing Rule-Based and LLM-Based Methods to Enable Active Robot Assistant Conversations" (2024) CUI@CHI 2024.

[21] Cf Nicholas Carlini et al. (n 15) 2634.

[22] Sascha Heyer, "Generative AI – Mastering the Language Models Parameters for Better Outputs" (2023) <https://medium.com/google-cloud/generative-ai-mastering-the-language-model-parameters-for-better-outputs-a82b07b4e383> accessed 10 December 2024.

[23] The term "hallucination" is misleading insofar as LLMs do not perceive anything – they rather confabulate, Beren Millidge, "LLMs Confabulate Not Hallucinate" (2023). <https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/> accessed 10 December 2024.

[24] Michael Schade, "How ChatGPT and Our Language Models Are Developed, Is Personal Information Used to Teach ChatGPT?" (2023) <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed> accessed 10 December 2024.

techniques – especially based on LLMs[25] – might allow for anonymising large amounts of text in an automated matter while preserving all relevant information. What results can be achieved with completely fictional[26] or synthetic data is an open research question and depends on the intended use cases.

It is important to distinguish training LLMs with personal data on the one hand and prompting them with personal data on the other hand. When prompted with personal data (for example, "Who is [name]?"), models usually generate outputs that include personal data (for example, "[name] is known for her expertise in the field of computer science"). However, most existing LLMs are not self-learning, ie, they do not use information from a prompt in other sessions. By contrast, during training, the models "memorise" information from the training data, ie, that information is reproducibly stored in the model parameters. For terms and texts that are prevalent in a large subset of the training data this happens intentionally. This ensures, for example, that LLMs use common terms such as "artificial intelligence" rather than creating strange terms such as "artificial smartness", and that they can provide information about historic events. "Memorisation", however, also occurs unintentionally for pieces of texts that are prevalent only in a small subset of training data.[27] Albeit this phenomenon has not been extensively researched yet, experiments with existing models indicate that duplicates in the training data promote unintentional "memorisation",[28] and that larger models with more parameters memorise more information from the training data than smaller ones.[29] In one experiment with GPT-2, the predecessor of the model that ChatGPT is based on, researchers were able to reproduce personal contact information from the training data. For ChatGPT, researchers could develop an attack to extract training data in a scalable matter.[30] It is likely that further breakthroughs in LLM research involve more efficient methods to extract training data. When addressing memorization, attention should be given not only to personal data but also to copyrighted material and trade secrets within the training data.

## III. Potential use cases for LLMs in the context of administrative decision-making

When discussing potential use cases for LLMs[31] in the context of administrative decision-making, two perspectives should be considered. On the one hand, LLMs might prove useful in the decision-making process (section C.I). On the other hand, LLMs could provide new tools for concerned individuals (section C.II).

### 1. LLMs in the decision-making

LLMs are able to perform various minor tasks that can enhance decision-making processes, eg, they have been found useful for processing and summarising large amounts of texts, or

---

[25] On the anonymisation of images Luca Piano et al., "Latent Diffusion Models for Attribute-Preserving Image Anonymization" (2024) arXiv:2403.14790.

[26] An LLM trained on completely fictional data could certainly not provide real-world facts.

[27] Carlini et al. (n 15) 2633.

[28] Nicholas Carlini et al., "Quantifying Memorization Across Neural Language Models" ICLR 2023 arXiv:2202.07646, 5; Nikhil Kandpal, Eric Wallace and Colin Raffel, "Deduplicating Training Data Mitigates Privacy Risks in Language Models" (39th International Conference on Machine Learning 2022). However, models also memorise information prevalent in small subsets of deduplicated training data, Carlini et al. (n 15) 2644.

[29] Carlini et al. (n 28) 4; Carlini et al. (n 15) 2645.

[30] Milad Nasr et al., "Scalable Extraction of Training Data from (Production) Language Models" (2023), arXiv:2311.17035.

[31] As stated above (section II.2 last para), LLMs refer to generative LLMs.

for generating ideas or multiple potential solutions for a specific problem.[32] However, LLMs could also play a major role in decision-making, particularly as they allow for the development of decision-support systems that draft complete decisions. LLMs can create texts with a high variance. In the context of administrative decisions, this makes them a promising tool for the automated generation of complex individual explanations that are not merely based on text blocks.

Taking an application for a building permit as an example, a simple LLM-assisted decision-making process could look as follows:

(1) The applicant files an application for a building permit with the competent authority and hands in all required documents electronically and in a standardised format.
(2) The information from the application is fed into an LLM-based decision-support system.
(3) The LLM drafts a complete decision including an explanation, and the LLM-based decision-support system provides the competent decision-maker with this draft.
(4) The competent human decision-maker checks the decision and issues either a decision as proposed by the decision-support system, or a slightly altered decision, or a completely different decision.

In step 3, the decision-support system could either fully rely on an LLM, or combine an LLM with another component or other components.[33] In the latter alternative, the other component would determine the decision to issue or deny a building permit, and the LLM would be prompted to generate the decision text including the explanation for this given decision. In the former alternative, the LLM would freely generate the decision text and, by this, determine the decision to issue or deny a building permit. Both alternatives require a domain-specific LLM that has been trained on specific training data, comprising not only relevant legal provisions, legal commentaries, documents on the legislative process and court rulings on these provisions, but also examples of decisions. As it has not been clarified yet whether, and if so, under which conditions, training on synthetic data could achieve satisfying results and the effective anonymisation of training data is not yet possible with reasonable effort,[34] the training would have to rely on historical real-world decisions. Where there are not enough real-world decisions available to train a performant model, LLM-based data augmentation technologies can be applied to expand and diversify the training data set by creating variations of available data in an automated manner.[35]

In step 4, the decision draft of the LLM serves as the starting point for the human decision-maker's considerations. The human decision-maker's task consists in thoroughly assessing the suggested decision rather than developing their own decision from scratch or composing it from existing text blocks and, usually, additional text for the individual case. Unlike text blocks, the draft of the LLM already consists of a complete and continuous text that has been generated for the specific case. Compared to writing a decision from scratch or composing it based on text blocks, taking an LLM's draft as a starting point of administrative decision-making has the potential to largely reduce the time that human decision-makers take to extract decision-relevant information from complex case files and arrive at their decision.

[32] Eva Eigner and Thorsten Händler "Determinants of LLM-Assisted Decision-Making" (2024) arXiv:2402.17385v1.

[33] For example, an LLM could be combined with a rule-based system, see Leismann, Wang and Mayer (n 20).

[34] See section II.3.

[35] Bosheng Ding et al., "Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges" (2024) arXiv:2403.02990.

LLMs could draft various other kinds of administrative decisions, eg, the decision of a data protection authority to impose an administrative fine on a data controller under the GDPR[36],[37] the decision of an authority to ban certain slogans at a demonstration, or a financial supervisory authorities' decision to deny an applicant the authorisation to provide financial services. LLMs could especially facilitate the efficiency of administrative decision-making where this requires processing large amounts of texts that take substantial efforts to read, analyse or write, ie, where the decision-making is based on comprehensive file material, or the decision requires extensive explanations. Furthermore, LLMs are capable of generating highly varied texts that cannot be composed from commonly used text modules, making them especially valuable in unique cases that demand a thorough individual evaluation.

### 2. LLMs as a tool for concerned individuals

Potential use cases for LLMs in the context of administrative decisions are not limited to the decision-making but the models could also prove useful for addressees of administrative decisions. LLMs could classify, translate, or summarise decisions albeit these tasks require especially high precision in the legal context. LLMs furthermore could provide in-depth explanations of decisions and possible responses, legal analyses to identify legal errors, and even draft complaints. Such tools could either support legal counsels of concerned individuals. They could make legal remedies available to especially vulnerable groups that lack access to legal counsels[38] by generating complaints that concerned individuals can directly file with the competent authority or court. A hackathon on generative AI held in June 2023 aimed to explore, among other things, tools based on LLMs to translate legislation and other legal documents into simpler language,[39] or to generate human right appeals for asylum seekers in an automated manner.[40]

### IV. Selected practical and legal challenges

Whether or not LLMs enable sufficiently good tools for both decision-makers and individuals who are concerned by decisions is an open research question. There are various concerns and doubts about both the practical feasibility of such tools and their legal compliance. The following sections introduce some of these challenges and, where applicable, highlight connections between them. Concretely, the article introduces practical and legal challenges regarding the suitability of LLMs for legal reasoning tasks (section IV.1), human oversight over LLMs (section IV.2) and the "memorisation" of training data (section IV.3).

### 1. Suitability of LLMs for legal reasoning tasks

Making, assessing or contesting an administrative decision involves solving legal reasoning tasks. Concretely, it is necessary to correctly apply legal norms to a specific

---

[36] Art 83 GDPR.

[37] See also example in section IV.1.b.

[38] See also Jens Frankenreiter and Julian Nyarko, "Natural Language Processing In Legal Tech" in David Engstrom (ed), *Legal Tech and the Future of Civil Justice* (Cambridge, Cambridge University Press 2023) 70 [76] <https://doi.org/10.1017/9781009255301.005> accessed 10 December 2024, who expect natural language processing techniques to become especially relevant for those who cannot afford a lawyer.

[39] Cf also Frankenreiter and Nyarko (n 38) 76.

[40] Hack to the Rescue, Challenges (2023) via Wayback Machine <https://web.archive.org/web/20231202225502/https://hacktotherescue.org/generativeai/challenges> accessed 10 December 2024, see the challenges proposed by Demokracja Przyszlosci, Socialbee and Asylex in particular.

case to draft, assess or contest an administrative decision. Legal reasoning can be highly complex as numerous legal norms can apply to a case and their application regularly involves interpreting what their often-ambiguous wording means.[41] Legal reasoning requires considering all relevant circumstances of a case. Administrative bodies are legally obliged to do so. On the EU level, the general principle of good (or "sound") administration[42] requires authorities to carefully examine all relevant facts of an individual case[43] and investigate further where necessary. This requirement can also be derived from the right to good administration codified in Article 41 CFREU.[44] According requirements can be found in Member State Law. For example, § 24 paras 1, 2 of the German Federal Administrative Procedure Act (Verwaltungsverfahrensgesetz) explicitly require authorities to take into account all circumstances relevant to the individual case, and to investigate the facts *ex officio*. The law regularly grants administrative bodies discretion[45] in decision-making, meaning there can be multiple lawful outcomes to a case. However, for legal reasoning tasks, it is essential to ensure consistency that is also regarded a general principle in EU administrative law[46]: On the one hand, an administrative decision, analysis or complaint must be consistent in itself, ie, especially not self-contradictory, on the other hand, administrative decisions must be consistent with other decisions to ensure compliance with principles of equal treatment, as laid down in Article 20 CFREU or Article 3(1) of the German Basic Law (Grundgesetz). Ensuring equality requires that similar situations are not treated differently and different situations are not treated identically without an objective justification.[47] However, the administration can change decision-making practices in the light of new technological, legal or societal developments. For example, new studies on training data extraction[48] can demand changes to a data protection authorities decision-making practice concerning the lawfulness of data processing with LLMs.[49]

Legal reasoning is not limited to finding a (or the) lawful outcome (eg, grant/deny a building permit) but also includes providing its justification, i.e. a clear explanation of the decision. Administrative bodies are obliged by law to provide reasons for their decisions. This requirement can be found in Member States' administrative procedural laws such as § 39 of the German Federal Administrative Procedure Act[50] that requires written or electronic administrative acts to be accompanied by a statement of grounds.[51]

---

[41] Parycek, Schmid and Novak (n 20) 8401.

[42] Cf CJEU Case C-531/12 P *Commune de Millau*, para 97; CJEU Case C-625/15 P *Schniga GmbH*, para 47.

[43] CJEU Case C-691/15 P *Bilbaína de Alquitranes*, para 35.

[44] Paul Craig, in Steve Peers, Tamara Hervey, Jeff Kenner and Angela Ward (eds), *The EU Charter of Fundamental Rights: A Commentary* (2021) Art 41 para 41.25 with reference to CJEU Case C-269/90 *Technische Universität München*, para 14.

[45] On discretion see Saar Alon-Barkat and Madalina Busuioc, "Human-AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice" 2023 (33:1) JPART 153, 154; Parycek, Schmid and Novak (n 20) 8402; Kate Vredenburgh, "AI and Bureaucratic Discretion' (2023) Inquiry <https://doi.org/10.1080/0020174X.2023.2261468> accessed 10 December 2024.

[46] Cf Diana-Urania Galetta et al., "The General Principles of EU Administrative Procedural Law – In-depth analysis for the JURI Committee" (Brussels, European Parliament 2015) 10 <https://op.europa.eu/en/publication-detail/-/publication/e1c335bd-f4c1-4595-8246-30fcfb8c3543> accessed 10 December 2024.

[47] On the general principle of equality CJEU Case C-217/91 *Spain v Commission* Para 37. Also cf Christopher McCrudden and Sacha Prechal, "The Concepts of Equality and Non-Discrimination in Europe: A Practical Approach" (2009) <https://ec.europa.eu/social/BlobServlet?docId=4553> accessed 10 December 2024.

[48] See sections II.3 and IV.3 on memorisation.

[49] See section IV.3. with n 130.

[50] Administrative Procedure Act [Verwaltungsverfahrensgesetz (VwVfG)] of 25 May 1976.

[51] On Belgian, Dutch and French Law see Ingrid Opdebeek, Stéphanie De Somer "The Duty to Give Reasons in the European Legal Area – A Mechanism for Transparent and Accountable Administrative Decision- Making? A Comparison of Belgian, Dutch, French and EU Administrative Law' 2016 (2) RAP 97.

Furthermore, in EU administrative law, a duty to give reasons follows from Article 41(2)(c) CFREU, Article 296(2) TFEU, and as a general principle of law.[52] The main rationales of the obligation are the provision of information to the concerned individual on the one hand, and enabling courts and administrative bodies to review decisions on the other hand.[53] To enable concerned individuals to understand and challenge decisions concerning them and courts to exercise their power of review, administrative bodies must "disclose in a clear and unequivocal fashion [their] reasoning",[54] stating in a concise and understandable manner all relevant facts, legal norms and decisive legal considerations.[55] The duty to give reasons not only ensures transparency but also facilitates a careful assessment of the case by decision-makers. The extent of the duty to give reasons, however, depends on the decision, and the requirements are particularly low for decisions that are in line with a consistent decision-making practice.[56]

With respect to legal reasoning, the use of LLMs to draft, analyse or contest administrative decisions poses two main research questions:

(1) *Feasibility of legal drafting with LLMs*: Are LLMs fit to solve the legal reasoning tasks outlined in section III, ie, can they generate complete decision drafts, analyses or complaints that are consistent with the facts resulting from the case file and the applicable legal norms?

(2) *LLMs and the duty to give reasons*: Can administrative bodies who issue decisions that are drafted by LLMs fulfil their duty to give reasons and, if so, how?

In the context of research question (1), it is questionable whether and, if so, to which extent legal reasoning can rely on language in general (section IV.1.a) and on LLMs in specific (section IV.1.b). Furthermore, LLMs in the context of administrative decision-making must meet consistency requirements but also allow for changes of administrative practices (section IV.1.c). For LLMs to prove useful for legal drafting in the context of administrative decisions, errors and biases must be avoided and robustness against manipulation must be ensured (section IV.1.d). Section IV.1.e addresses research question (2).

### a. Language as a limited medium of communication

Some experts point out that LLMs, in general, have limited abilities as they are based on language which is a limited medium of communication.[57] At first glance, this concern does not seem to apply to legal tech use cases as the law itself highly relies on language. However, legal language is particularly difficult as it is charged with ambiguous terms that are often inconsistently used in different legal acts and that therefore require careful interpretation not only based on the mere wording but also in the light of the specific relationships between legal provisions, their historical context, and their intent and purpose. For example, determining whether a controller can claim legitimate interest according to Article 6(1)(f) GDPR requires a thorough risk assessment and balancing of

---

[52] Fink and Finck (n 10) with further references. On the relation between the rights under the CFREU and unwritten principles of EU law Herwig C.H. Hofmann and Bucura C. Mihaescu "The Relation between the Charter's Fundamental Rights and the Unwritten General Principles of EU Law: Good Administration as the Test Case" 2013 (9) ECLR 73.

[53] Cf CJEU Case C-22/94 *Irish Farmers Association and Others*, para 39; Case 2/56 *Geitling*, Summary para 2.

[54] On Art 253 EC, the predecessor provision of Art 296(2) TFEU CJEU Case C-113/00 *Kingdom of Spain* para 47.

[55] Fink and Finck (n 10) 383.

[56] Cf Case C-228/99 *Silos e Mangimi Martini SpA v Ministero delle Finanze* paras 27 ff.

[57] Jacob Browning and Yann Lecun, "AI And The Limits of Language" (2022) <https://www.noemamag.com/ai-and-the-limits-of-language/> accessed 10 December 2024.

interests. To give another example, the term "Eigentum" that can be translated to property, in German civil law, refers to the ownership of tangible objects only,[58] while the constitution[59] uses the term in a broader meaning, referring also to the ownership of rights. Legal laypersons using legal terms without knowing their exact legal meaning adds to the problem. LLMs can only prove useful for drafting, analysing and contesting administrative decisions if it is possible to engrain legal semantics in their parameters. Whether the models allow for that with the necessary precision, and, if so, which effort it takes to train and fine-tune such models, is an open research question that can be answered only through experiments and empirical studies with domain-specific LLMs, as results of attempts to train LLMs for legal use cases[60] and studies on existing general-purpose models[61] do not generalise across domains and to sufficiently fine-tuned models.

### b. Limited reasoning abilities of LLMs

From a pessimistic point of view, it could be assumed that LLMs are simply unfit for making, analysing and contesting administrative decisions as existing models have shown limited reasoning abilities.[62] Indeed, there are limited studies to solve legal problems with public general-purpose models and fine-tuned models[63] that are largely phenomenological, limited to specific applications and models and do not permit drawing general

---

[58] §§ 903, 90 of the German Civil Code (Bürgerliches Gesetzbuch, BGB).

[59] Art 14 of the German Basic Law (Grundgesetz, GG).

[60] See especially Liga and Robaldo (n 5); Ilias Chalkidis et al., "LEGAL-BERT: The Muppets Straight Out of Law School" (2020) arXiv:2010.02559; Shunyu Yao et al., "Lawyer GPT: A Legal Large Language Model with Enhanced Domain Knowledge and Reasoning Capabilities" in *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering (RAIIE '24)* 108.

[61] See for example Jonathan J Choi and Daniel Schwarcz, "AI Assistance in Legal Analysis: An Empirical Study' 73 Journal of Legal Education (forthcoming 2024) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4539836> accessed 10 December 2024; Lauren Martin et al., "Better Call GPT, Comparing Large Language Models Against Lawyers" (2024) arXiv:2401.16212; John J Nay et al., "Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence" 2024 (382) Philosophical Transactions of the Royal Society 20230159.

[62] Seungpill Lee et al., "Reasoning Abilities of Large Language Models: In-Depth Analysis on the Abstraction and Reasoning Corpus" (2024) arXiv:2403.11793. On Measuring Legal Reasoning in LLMs Neel Guha et al., "LEGALBENCH: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models" (2023) NeurIPS. Also cf Morison and McInerney (n 19) 26 f.

[63] See for example Ilias Chalkidis, "ChatGPT may Pass the Bar Exam soon, but has a Long Way to go for the LexGLUE Benchmark" (2023) arXiv:2304.12202; Chalkidis et al. (n 60); Wentao Deng et al., "Syllogistic Reasoning for Legal Judgment Analysis" in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 13997; Neel Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models" (2023) arXiv:2308.11462; Cong Jiang and Xiaolei Yang, "Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction" arXiv:2307.08321; Xiaoxi Kang et al., "Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?" in *Findings of the Association for Computational Linguistics: EMNLP* 2023, 13900 (Singapore, Association for Computational Linguistics); Daniel Martin Katz et al., "GPT-4 Passes the Bar Exam" (2024) 382 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 20230254; Dietrich Trautmann, "Large Language Model Prompt Chaining for Long Legal Document Classification" (2023) arXiv:2308.04138; Dietrich Trautmann, Alina Petrova and Frank Schilder, "Legal Prompt Engineering for Multilingual Legal Judgement Prediction" (2022) arXiv:2212.02199; Shaurya Vats, 'LLMs – the Good, the Bad or the Indispensable?: A Use Case on Legal Statute Prediction and Legal Judgment Prediction on Indian Court Cases" in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12451 (Singapore, Association for Computational Linguistics); Fangyi Yu, Lee Quartey and Frank Schilder, "Legal Prompting: Teaching a Language Model to Think Like a Lawyer" (2022) arXiv:2212.01326; Fangyi Yu, Lee Quartey and Frank Schilder, "Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks" *Findings of the Association for Computational Linguistics: ACL* 2023, 1358 (Toronto, Canada, Association for Computational Linguistics). For a broader overview on legal argumentation based on natural language processing see Santosh T Y S S et al., "Towards Supporting Legal Argumentation with NLP: Is more Data Really All You Need?" (2024) arXiv:2406.10974v1 with further references.

conclusions.[64] What makes legal reasoning through LLMs especially challenging is that LLMs rely on statistical relationships, ie, correlations, while law relies on causalities[65]:

Taking a decision of a data protection authority to impose an administrative fine on a controller as an example, in terms of causality, the fine is imposed

  – *because* the controller's data processing is not compliant with the GDPR (Article 83 GDPR)
   – *because* they lack a legal basis (Article 6(1) GDPR)
    – *because* they obtained unlawful consent
     – *because* the data subject has not unambiguously indicated their wish to consent (Article 4(11) GDPR)
      – *because* the controller obtained consent through a pre-ticked-box (cf recital 31 GDPR)
    – *and because* no other legal basis applies
  – *and because* the fine is effective, proportionate and dissuasive (Article 83(1), (2) GDPR).

By contrast, a – correlation-driven – LLM could, for example, generate an according decision including the phrase "In accordance with Article 58(2)(i) and Article 83 GDPR, the [specific supervisory authority] hereby imposes an administrative fine" *based on* the prevalence of this phrase in decision samples within the training data that also contain the word combination "pre-ticked box" and other correlated elements. However, as existing LLMs are opaque machine learning models,[66] it is very difficult and more often impossible to map elements of the generated output text to specific elements of the prompt.

Albeit LLMs do not follow the logic of the law and reach their results fundamentally different from humans, it seems premature to conclude LLMs cannot assist in (legal) reasoning tasks per se.[67] An LLM can prove useful for the use cases outlined above if it just mimics or simulates legal reasoning[68] well enough. For example, think of a hypothetical highly performant LLM that drafts a lawful decision according to which a building permit is denied due to an insufficient setback distance. Even though that LLM had no understanding of the meaning of words, the output of such a model could be based on relevant information in the prompt, namely the part of the text in the application for the building permit which contains the information that allows for calculating the distances of the building to neighbouring buildings, property lines, streets, etc. However, a particular challenge lies in new legal norms for which there are no historical case data to train models on, or new judgments that require changes to decision-making practices.[69] Bringing together IT experts and experienced administrative professionals is imperative to build and use performant models.

Ongoing research explores various approaches to enhance LLMs' performance in reasoning tasks. For existing models, researchers have proposed to solve legal problems with LLMs through prompt engineering, especially through breaking down prompts to series of reasoning steps.[70] GPT-4 can solve analytic problems by generating Python code

---

[64] Pessimistic Frankenreiter and Nyarko (n 38) 82 ff.

[65] Referring to AI models in general (Fink and Finck (n 10) 385.

[66] See section II.1.

[67] See for example Subbarao Kambhampati, "Can Large Language Models Reason and Plan?" (2024) 1534 Annals New York Academy Science 15. Frankenreiter and Nyarko (n 38) [84, 90] argue natural language processing will remain irrelevant for predicting legal outcomes and limited to information extraction tasks.

[68] Cf Morison and McInerney (n 19) 6.

[69] Cf Santosh T Y S S et al. (n 63) 7 with technical approaches to solve that problem and with further references.

[70] Jiang and Yang (n 63); Yifei Li et al., "Making Large Language Models Better Reasoners with Step-Aware Verifier" (2023) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* Vol 1, 5315 (Toronto, Canada, Association for Computational Linguistics); Trautmann (n 63); Trautmann, Petrova and Schilder

that simulates the task.[71] In experiments with GPT-4 comprising various tasks from various domains, the model has been found to solve various tasks that require reasoning.[72] Other researchers fine-tune LLMs on logical reasoning datasets,[73] or combine LLMs with symbolic solvers[74] to enable them to perform complex reasoning tasks. It remains an open research question, however, to which extent the observations in studies with existing general-purpose LLMs generalise across domains and to sufficiently fine-tuned models.

### c. Consistency and changing decision-making practices

In terms of consistency, public general-purpose models have cast doubt about the usefulness of the models for tasks that require consistency. Existing public general-purpose LLMs such as ChatGPT-4 or Gemini usually generate inconsistent outputs, ie, when prompted with the same text prompt multiple times, they will generate different outputs each time. This not only concerns the mere wording (eg, data processing "lacks a legal basis"/"is not lawful"[75]), the variance of which may be irrelevant or even desirable for legal applications, but also the meaning of the output (eg, data processing is unlawful/based on consent). If an LLM generates two significantly different outputs for identical inputs, it is probably also more likely to produce divergent outputs for merely similar prompts, which could render an administrative decision unlawful by violating the principle of equal treatment.[76] For example, if an LLM generates two conflicting decisions – one granting a building permit and the other denying it – for the same case file, it is likely to generate inconsistent decisions for two only slightly similar cases. LLMs are not output-inconsistent per se. After an update to the GPT API,[77] it is possible to generate more consistent outputs with GPT-4.[78] However, assessing LLMs' usefulness in the context of administrative decision-making requires testing whether the models can meet the specific consistency requirements for legal use cases – namely, self-consistency in their outputs and, in administrative decision-making, alignment with the decision-making practices of the respective administrative body.

Albeit consistency in administrative decision-making is desirable and necessary in general, sometimes administrations need and aim at change. For example, a building authority might want to set a new course for the future and change certain decision-making practices, eg, to grant building permits in cases where they would previously have been denied. One concern about the use of machine learning models in administrative decision-making is that the models are trained on historical data and might rather "predict the past" than imagine a future.[79] LLMs seem to solve this problem to some extent as they

---

(n 63); Yu, Quartey and Schilder (2022) (n 63); Yu, Quartey and Schilder (2023) (n 63); Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" (2023) arXiv:2201.11903v6.

[71] Russell A Poldrack, Thomas Lu and Gašper Beguš, "AI-Assisted Coding: Experiments With GPT-4" (2023) arXiv: 2304.13187v1.

[72] Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early experiments with GPT-4" (2023) arXiv: 2303.12712.

[73] Yanda Li et al., "Reason from Fallacy: Enhancing Large Language Models" (2024) Logical Reasoning through Logical Fallacy Understanding arXiv:2404.04293v1.

[74] See, for example, Marius Constantin Dinu, "SymbolicAI: A Framework for Logic-Based Approaches Combining Generative Models and Solvers" (2024) arXiv:2402.00854; Liangming Pan, "LOGIC-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning" (2023) arXiv:2305.12295.

[75] Cf Art 6 GDPR.

[76] See section IV.1 first para.

[77] Application Programming Interface. The GPT API enables developers to integrate OpenAI's models into their own services, see OpenAI, API <https://platform.openai.com/> accessed 10 December 2024.

[78] Jordan Gibbs, "GPT Seed Selection is a Game Changer for Consistent & Fair AI Classification" (2023) <https://medium.com/@jordan_gibbs/gpt-seed-selection-is-a-game-changer-for-consistent-fair-ai-classification-5200635b54da> accessed 10 December 2024.

[79] Cf Moritz Hardt and Michael P Kim, "Is Your Model Predicting the Past?" (2023) EAAMO Article No 5.

can generate novel outputs with high variance rather than being determined by the training data.[80] This might make decision-support systems based on LLMs less prone to perpetuating decision-making practices from the past than systems based on other models. However, administrative bodies typically plan changes to their decision-making practices with a specific vision for the future in mind. While an LLM could draft a vision for the future of an administrative body and propose concrete measures to work towards this future, existing LLMs have shown poor planning capabilities.[81] However, the research on LLMs abilities is still at an early stage, and findings from experiments on existing LLMs, at best, provide little information on other, especially domain-specific models. When authorities implement decision-support systems based on LLMs it will be crucial to ensure that the decision-making is sufficiently consistent and changes of decision-making practices do not happen randomly but planned, regardless of whether the LLM is used for planning or not.

### d. Errors, bias and manipulation

A main concern about the automation of making, analysing and contesting decisions based on LLMs is that the models may generate erroneous outputs such as an unlawful building permit, a flawed legal analysis of an administrative decision, or an unfounded complaint. Just as human decision-makers, any automated system, regardless of whether it is based on an AI model or a transparent algorithm, produces errors. Errors can render an administrative decision unlawful or a complaint against an administrative decision unsuccessful. While human decision-makers from the administration or legal counsels would likely be able to find and correct all errors, concerned individuals that fully rely on an erroneous LLM might misunderstand decisions concerning them or even file flawed complaints – with the risk of losing their legal remedies. From both a fundamental rights and an efficiency perspective, the use of decision-support systems that produce too many errors is not justifiable. Even decision-support systems that produce less errors than human decision-makers can violate principles of non-discrimination[82] if their errors disproportionally concern certain groups of individuals. There are many potential causes for erroneous decisions, for example, misinterpretations of legal terms,[83] poor reasoning abilities,[84] misinformation in the training data, training data information that is stored incorrectly or incompletely in the model parameters, or – for LLMs specifically – "hallucinations"[85] such as basing decisions on non-existing or non-applicable legal norms, court rulings,[86] or an incorrect summary of the case.

Furthermore, albeit LLMs' can generate outputs that are less determined by the training data,[87] LLMs – just as other machine learning models[88] – have been shown to adapt biases from the training data.[89] In the context of administrative decision-making, for example, a

---

[80] See section II.2.

[81] Sébastien Bubeck et al. (n 72); Karthik Valmeekam et al., "On the Planning Abilities of Large Language Models: A Critical Investigation" (2023) NeurIPS; Karthik Valmeekam et al., 'PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change" (2023) NeurIPS.

[82] See only Art 21 CFREU.

[83] See section IV.1 first para.

[84] See section IV.1.b.

[85] See section II.2.

[86] Molly Bohannon, "Lawyer Used ChatGPT In Court – And Cited Fake Cases. A Judge Is Considering Sanctions" (2023) <https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/> accessed 10 December 2024.

[87] See section II.2.

[88] On bias in artificial intelligence European Agency for Fundamental Rights (FRA), "Bias In Algorithms – Artificial Intelligence and Discrimination" (2023) <https://fra.europa.eu/en/publication/2022/bias-algorithm> accessed 10 December 2024.

[89] On LLMs Jessica Echterhoff et al., "Cognitive Bias in High-Stakes Decision-Making with LLMs" (2024) arXiv: 2403.00811v1; Masahiro Kaneko et al., "Evaluating Gender Bias in Large Language Models via Chain-of-Thought

model could be biased against Black applicants, when trained with historic decisions in which building permits by applicants from a predominantly Black neighbourhood have been disproportionately denied. In experiments with existing LLMs, biased outputs could be avoided by telling the model not to be biased (in general or with more specific instructions).[90] Another study led to the conclusion that LLMs are capable of "moral self-correction" from a size of 22 billion model parameters and improve in that task with increasing model size and reinforcement training based on human feedback.[91,92] However, bigger models can cause other problems, for example, they can facilitate training data "memorisation".[93]

Also, malicious attacks[94] can cause incorrect outputs. For example, researchers were able to compromise existing LLMs' outputs through malicious instructions in prompts (prompt injection).[95] Robustness against malicious attacks is especially important in administrative decision-making: Who applies for a building permit, must not be able to trick the system into issuing an unlawful building permit by designing the application in a certain way.

When exploring LLMs in the context of administrative decision-making, mitigation techniques[96] should be implemented and tested for specific use cases and user groups. Sufficiently checking LLMs for hidden bias and vulnerabilities might be possible through extensive prompting with variations of inputs[97]: A model could, for example, be fed with slightly varied applications for building permits that do or do not contain discriminatory factors or information that serves as proxy for such information,[98] ie, indirectly refers to it (eg, an address or a name that indicate the ethnic background of the applicant). Through comparing the outputs of the model then, it can be possible to identify bias, for example if, for the same case the model issues a building permit, if the applicant has a German name, but denies it if the name in the application is changed to an Arabic name. Such tests could also be employed to investigate cases where individuals challenge a decision affecting them as discriminatory.

### e. LLMs and the duty to give reasons
Even if an LLM drafts a lawful decision with reasons that justify the decision and are compliant with the applicable legal norms, it is questionable whether authorities that issue

---

Prompting" (2024) arXiv:2401.15585v1; Roberto Navigli, Simone Conia and Björn Ross, "Biases in Large Language Models: Origins, Inventory, and Discussion" 2023 (15) Journal of Data and Information Quality no 10. On image generative models Mi Zhou et al., "Bias in Generative AI" (2024) arXiv:2403.02726.

[90] Ganguli et al., "The Capacity for Moral Self-Correction in Large Language Models" (2023) arXiv:2302.07459v2.

[91] See section II.1 above.

[92] Ganguli et al., "The Capacity for Moral Self-Correction in Large Language Models" (2023) arXiv:2302.07459v2.

[93] See sections II.3 and IV.7.

[94] Erfan Shayegani et al., "Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks" (2023) arXiv:2310.10844.

[95] Jinwei Yi et al., "Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models" (2024) arXiv:2312.14197; Yi Liu et al., "Prompt Injection Attack against LLM-Integrated Applications" (2024) arXiv:2306.05499v2.

[96] On hallucination mitigation S M Towhidul Islam Tonmoy et al., "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models" (2024) arXiv:2401.01313.

[97] Finale Doshi-Velez et al., "Accountability of AI Under the Law: The Role of Explanation" revised (2019) arXiv:1711.01134 14.

[98] In the context of special categories of personal data according to Art 9 (1) GDPR CJEU Case C-184/20 *Vyriausioji tarnybinės etikos komisija* paras 117 ff. Also see Finale Doshi-Velez et al., "Accountability of AI Under the Law: The Role of Explanation" revised (2019) 14 f. Yannick Meneceur, "Artificial Intelligence, Public Administration, and the Rule of Law – Towards the Uncertainties of a New 'Rule of Algorithm'" in Markku Suksi (ed), *The Rule of Law and Automated Decision-Making* (2023) 117, 132 (Cham, Springer Nature Switzerland AG) points out that the total elimination of bias from a training dataset is not realistic.

this decision fulfil their duty to give reasons (research question (2)). This is especially problematic insofar as existing LLMs are based merely on statistic relationships between words and can only mimic an understanding of their meaning[99] and of legal reasoning.[100,101] They draft their outputs word for word, predicting based on the statistical distributions in the training data which word follows another in a specific context – ie, the prompt and the text that the LLM has already generated for this prompt. Even a correct output, for example the lawful decision to deny a building permit, can be based on completely irrelevant elements from the prompt,[102] eg, the applicant's address or name when the model has been trained on historic case files including rejecting decisions addressing persons from the same street or with the same name. The reasons stated in a decision that has been generated by an LLM always differ from the reasons for which the LLM has generated this decision.

It is questionable what the duty to give reasons requires for fully or partially automated administrative decisions.[103] For machine-learning systems it has been argued that some degree of explainability is a prerequisite for fulfilling the duty to give reasons as the duty to give reasons requires explaining why a decision was reached and thus disclosing how different factors were weighed by the system.[104] Other legal scholars also deem explainability necessary for lawful AI systems,[105] even though there are trade-offs between model explainability on the one hand and model performance, usability and scalability and training data privacy on the other hand.[106]

However, it is crucial to consider the specific kind of decision-support that LLMs could enable. An LLM does not merely output a binary recommendation (eg, grant/deny a building permit) or a risk scoring (eg, a security risk posed by a third country national that wishes to enter the EU[107]), but a complete decision text. If an LLM is used for administrative decision-drafting, two cases must be distinguished: Where the decision generated by an LLM is issued without meaningful human involvement, the reasons stated in the generated text cannot satisfy the duty to give reasons, given the intent and purpose of this obligation. This is because it aims to ensure transparency of the decision-making and, by this, enable individuals, courts and administrative bodies to understand the decision and assess its lawfulness.[108] LLMs may generate decisions whose reasoning appears transparent, but the reasons generated by the models do not reveal the factors relevant to the decision and their weight as LLMs merely mimic reasoning. However, when a decision-draft by an LLM serves as a starting point for human decision-making, the

---

[99] Cf sections II.1, IV.1.b above.

[100] Morison and McInerney (n 19) 6: "LLMs may be able to simulate legal reasoning but cannot exercise it."

[101] Morison and McInerney (n 19) 26 f.

[102] Cf Miles Turpin et al., "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting" (2023) arXiv:2305.04388v2.

[103] Fink and Finck (n 10) 383 ff.

[104] Fink and Finck (n 10) 383 f. Agreeing with this view: Davide Liga, "The Interplay Between Lawfulness and Explainability in the Automated Decision-Making of EU Administration" in Herwig C H Hofmann and Felix Pflücke (eds), *Governance of Automated Decision-Making and EU Law* (Oxford, OUP 2024) 252 <https://doi.org/10.1093/9780198919575.001.0001> accessed 10 December 2024.

[105] See for example Parycek, Schmid and Novak (n 20) 8404 ff. On the functions of explainability Philipp Hacker and Jan-Hendrik Passoth, "Varieties of AI Explanations Under the Law. From GDPR to the AIA and Beyond" in Holzinger et al. (eds), *xxAI – Beyond Explainable AI*, International Workshop Held in Conjunction with ICML 2020, 343, 344 ff.

[106] Liga (n 104) 247 ff.

[107] On automated decisions at the EU Border Paulina Pesch and Franziska Boehm, "Data Protection and Machine-Learning-Supported Decision-Making at the EU Border: ETIAS Profiling Under Scrutiny" Hofmann and Pflücke (n 104); Paulina Pesch, Diana Dimitrova and Franziska Boehm, "Data Protection and Machine-Learning-Supported Decision-Making at the EU Border: ETIAS Profiling Under Scrutiny" in Agnieszka Gryszczyńska et al. (eds), AFP 2022 LNCS 13279, 50.

[108] See section IV.1. first para.

situation is different, even in cases where the human decision-maker issues the decision without any changes to the LLM's draft. This is because a human decision-maker who checks an LLM-drafted decision adopts the reasons the LLM has generated to the extent they do not make changes to them (manually or through prompting the model again). In other words, human decision-makers can make LLM-generated reasoning their own. In this situation, neither the concerned individual nor courts or administrations need to understand how the LLM reached its output to assess the reasoning of the decision. Then the functioning of the LLM is not the reason for the decision but simply the reason for the decision being made faster. To substantially explain their decision the human decision-maker does not need to explain the functioning of the LLM.[109] LLMs, therefore, are especially promising for legal use cases. The fact that the outputs of LLMs are fully verifiable, regardless of the transparency of their parameters, supports the implementation of decision support systems based solely on LLMs, ie, without combining them with other machine learning models whose lack of transparency could limit the verifiability of decisions. Whether or not there is a right to explanation under the GDPR that requires more transparency of the statistical weights of such models when they are used for automated decision-making has yet to be clarified.[110] It is questionable that such information would facilitate the exercise of individual rights.[111]

It is important to consider, however, that highly performant LLMs might compromise the purpose of the duty to give reasons, which is to enable a thorough assessment by the human decision-maker. This is because it is simply tempting for human decision-makers to just issue an LLM-drafted decision without proper scrutiny rather than to take the time to carefully assess the case. Furthermore, LLMs that generate convincing reasons might reinforce automation bias, the human tendency to over-rely on computer-generated outputs.[112] Especially there is the risk that users confuse an explanation provided by an insufficiently performant model[113] with transparency of its functioning. In practice, it will be hard to distinguish cases in which human reasoning has taken place and cases where the decision was made, in fact, fully automated by the LLM.

## 2. Human oversight over LLMs

As LLM-based decision-support systems bear a particular risk of human decision-makers issuing decision drafts without scrutiny,[114] implementing LLMs in administrative decision-making calls for safeguarding human oversight. Human oversight is required under two legal provisions: Firstly, data protection law grants individuals the right not to be subjected to solely automated decisions with legal effects on them.[115] A decision based on automated processing is not solely based on automated means only if a human

---

[109] Different view for AI in general Fink and Finck (n 10) 383 ff.

[110] Cf on AI-based automated decision-making Fink and Finck (n 10) 388 f.; on automated decision-making in general Sandra Wachter, Brent Mittelstadt and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" 2017 (7) IDPL 76.

[111] Also cf Hacker and Passoth (n 105) 349, who argue that the duty to provide data subjects with information on the decision-making logic in cases of profiling (Artt 13(2)(f), 14(2)(g), 15(1)(h) GDPR) comprises information that serves the norms' purpose and intent to enable the data subjects to exercise their individual rights. Also see Lilian Lilian Edwards and Michael Veale, "Slave To The Algorithm? Why A 'Right To An Explanation' Is Probably Not The Remedy You Are Looking For" 2017 (16) DLTR 18.

[112] Parycek, Schmid and Novak (n 20) 8407 f; Hannah Ruschemeier, "The Problems of the Automation Bias in the Public Sector – A Legal Perspective" in *Proceedings of the Weizenbaum Conference 2023: AI, Big Data, Social Media, and People on the Move*, 1; Saar Alon-Barkat and Madalina Busuioc, "Human-AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice" 2023 (33) JPART 153.

[113] See sections IV.1–5.

[114] See section IV.1.e last para.

[115] Cf Art 22(1) GDPR, Art 24(1) EUDPR.

decision-maker is not merely formally but meaningfully involved in the decision-making.[116] This requires human decision-makers to effectively oversee decision-support systems, ie, to scrutinise their outputs. Secondly, for high-risk AI systems Article 14 of the AI Act[117] stipulates human oversight. Furthermore, as pointed out above,[118] it is necessary that human decision-makers scrutinise drafts generated by LLMs to fulfil their duty to provide reasons when issuing a decision with LLM-generated reasons.

When implementing LLMs in the context of administrative decisions, it is crucial to avoid automation bias by design, ie, to design the systems in a way that facilitates critical assessment.[119] Legal scholars assume that explainability[120] facilitates control and oversight over AI systems.[121] However, this view lacks differentiation. Explainable (or interpretable) AI can either help human decision-makers to identify erroneous outputs, or reinforce automation bias as an explanation can make the output seem even more reliable.[122] Other approaches to facilitate human oversight should be considered. For example, it is possible to design a decision support system in such a way that it sometimes incorporates errors in its outputs, and informs human decision-makers of this in order to undermine their trust in the system.[123] Any LLM-based decision-support system should be deployed only after thorough testing and empirical validation to ensure human oversight. This requires user studies with the specific systems, tailored to the specific use case (eg, decisions to grant/deny building permits) and involving members of the particular group of administrative decision-makers who are intended to use the system.[124] This is because, personal and contextual factors as well as tool characteristics can affect automation bias.[125] Furthermore, to enable human decision-makers to effectively oversee LLM-based decision-support systems, it is not only important that they have an understanding of the functioning and limitations of LLMs in general[126] (also cf Article 13(3)(b) AI Act), but also experiences with the specific system.[127]

Automation bias is not only a problem in the decision-making but also can impair the benefits of tools that are designed to help those who are concerned by decisions. When legal counsels or the concerned individuals themselves over-rely on the outputs of a LLM, this might result in individuals not effectively exercising their remedies or even losing them. Consequently, tools should only be released after representative user studies and be subject to regular audits as both decision contexts and users might change over time. This task is especially challenging for public tools that are intended to be used by legal laypersons from various educational and language backgrounds.

---

[116] Article 29 Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (2017) wp251rev.01 20 f.

[117] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence.

[118] See section IV.1.e.

[119] Cf Art 25 GDPR.

[120] On interpretability and explainability see Liga (n 104) with further references.

[121] Ben Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms" (2022) accessed 10 December 2024; Maia Jacobs et al., "How Machine-Learning Recommendations Influence Clinician Treatment Selection" 2021 Translational Psychiatry Article no. 8; Forough Poursabzi-Sangdeh et al., "Manipulating and Measuring Model Interpretability" (2021) arXiv:1802.07810.

[122] Poursabzi-Sangdeh et al. (n 121).

[123] Cf Marwa Gadala, "Automation Bias: Exploring Causal Mechanisms and Potential Mitigation Strategies" (2017) <https://openaccess.city.ac.uk/id/eprint/17889/> accessed 10 December 2024, 71.

[124] In the context of the European Travel Information and Authorisation System (ETIAS) Pesch, Dimitrova and Boehm (n 107) 60.

[125] Gadala (n 123) 35 ff.

[126] On blockchain forensics Michael Fröwis et al., "Safeguarding the Evidential Value of Forensic Cryptocurrency Investigations" 2020 (33) FSI: Digital Investigation 200902.

[127] In the context of LLM-based applicant selection Kätscher and Pesch (n 8) 51.

### 3. *"Memorisation" of training data*

Some of the biggest legal problems of existing LLMs arise from the "memorisation"[128] of training data. That information from the training data, during training, is reproducibly stored in the model parameters poses risks especially where this information comprises personal data. The training of an LLM in the context of administrative decision-making requires various data that also comprise personal data: Legislative materials contain the names of individuals involved in the legislation process. Court rulings include names of judges and persons involved in the case. Historic case files refer to applicants, addressees of decisions and other concerned individuals. Research articles and commentaries comprise information on authors. To the extent personal training data are consistently stored within a model's parameters in such a way that personal information can be extracted by prompting the model, both the LLM and its outputs, when such data is reproduced, constitute personal data.[129] The probability that memorised personal data are reproduced is arguably higher for publicly available LLMs. This is because a bigger group of potential users, including not only individuals that are subject to administrative decisions but also other interested groups such as researchers[130] and malicious attackers, might prompt the model to extract training data from it, or even accidentally cause the model to output training data. Where LLMs include memorised personal training data in decisions or complaints, users are at risk of unlawfully processing personal data, especially but not only where the data are inaccurate[131] due to "hallucinations". For example, an LLM could feed information on an addressee of a historic decision from the training data into a decision on a building permit.

Similarly, when "memorisation" concerns copyrighted material,[132] users might unintentionally infringe copyrights. For example, an LLM might include a section from a legal commentary in the explanation of a decision or a complaint without referencing the source. Furthermore, "memorisation" can concern business or trade secrets – with the risk that these secrets are disclosed to users of the system and shared with others. For example, an LLM used for decision-drafting in a building authority might feed information on a manufacturing process of a company into a decision concerning another company, and a human decision-maker might not remove this information from the draft before issuing the decision.

To develop decision-support systems and tools for concerned individuals that are compliant with data protection and copyright requirements, investigating "memorisation" is crucial. It is an open research question to which extent findings on existing models apply to other, especially domain-specific models. Avoiding the use of personal data[133] or copyrighted material in training entirely seems impractical and would impair a model's ability to include correct references in its outputs. Also, "memorisation" can be desirable in certain use cases for some texts from the training data.[134] To which extent "memorisation" or the reproduction of training data in the output should be avoided must be determined for the specific use case, model, user group, and training data, considering the risks of training data reproduction. In any case, it seems necessary to efficiently avoid the reproduction of personal training data and unreferenced copies of training texts. When addressing other challenges of LLMs in the context of administrative decision-making, it is

---

[128] See section II.3.

[129] Pesch and Böhme (n 8) 920 f. Also, the models may generate new personal data, ibid 921.

[130] On training data extractions studies see section II.3.

[131] On data accuracy in the context of LLMs Pesch and Böhme (n 8) 921 f.

[132] On LLMs and their outputs as copies under German copyright law Pesch and Böhme, "Artpocalypse Now? Generative KI und die Vervielfältigung von Trainingsbildern' (2023) GRUR 997.

[133] See section II.3 on anonymisation.

[134] See section II.3.

crucial to consider potential trade-offs with "memorisation" mitigation. For example, a model with more parameters might reach better results in legal reasoning tasks, but be more likely to reproducible storing training data information within its model parameters.

## V. Conclusion and outlook

Existing models' properties and abilities make LLMs a promising tool to increase the efficiency of administrative decision-making processes, and to help individuals to understand, analyse, and contest decisions concerning them. For human decision-makers in administrations, LLM-generated decision drafts could serve as a starting point for decision-making. However, the models pose various challenges that may impair both their performance and their compliance especially with requirements under administrative law, fundamental rights, data protection law, the AI Act and copyright law. A main challenge for LLMs in the context of administrative decision-making lies in training models that are sufficiently performant in legal reasoning tasks and thus able to generate complete legal assessments that are consistent with the specific case and applicable legal norms. There are numerous approaches to improve LLMs' performance in legal drafting. It is especially questionable whether LLMs need to be combined with rule-based systems that follow a formal logic, or LLMs can be trained to mimic legal reasoning with sufficient results based on mere correlations. It will be crucial to avoid errors, biased outputs and manipulation attacks. Ensuring model performance and accuracy is especially critical for LLMs that are intended to be used by legal laypersons such as concerned individuals that cannot afford a lawyer. When assessing LLMs' usefulness for administrative decision-making, their compatibility with the need for consistent administrative decisions on the one hand, and the need for planning changes in decision-making practices on the other hand must be considered. Unlike other opaque machine learning models, the outputs of LLMs can be fully examined for legal compliance. The article argues that the administration can, therefore, fulfil the duty to provide reasons even with LLM-generated reasons if a human thoroughly checks the LLM-generated draft. However, to facilitate a careful assessment of the case and to meet human oversight requirements, it is necessary to design administrative decision-support systems and tools for concerned individuals and their counsels in a way that facilitates human oversight. Which design elements and functionalities achieve this goal, is an open research question and depends on the concrete use case and user group.

Whether and if so, to which extent and under which circumstances, LLMs can facilitate making, analysing and contesting decisions, and how human oversight can be ensured, can only be assessed through fine-tuning domain-specific LLMs and testing them for specific use cases involving members of the specific user group. This is because observations from previous, inherently experimental research on existing models do hardly provide any insights about the performance and lawfulness of new, domain-specific LLMs in the context of administrative decision-making. The development of performant LLMs with the necessary safeguards requires a holistic view of requirements and challenges, as it is important to identify trade-offs, for example between bias prevention and "memorisation" prevention,[135] performance and "memorisation" prevention,[136] or explainability and performance.[137] This requires bringing together at least legal experts and practitioners to define legal and practical requirements, computer scientists to prepare training datasets, fine-tune domain-specific models, implement safeguards and run tests, and social scientists to carry out empirical user studies examining, in particular, user attitudes towards LLMs in the context of automation bias. Based on such studies, for specific use

---

[135] See section IV.1.d.
[136] See sections II.3 and IV.3.
[137] See section IV.1.b.

cases, it will be possible to determine the costs and benefits of the models' development and use, and to assess whether, and if so, to which extent LLMs can increase the efficiency of making, analysing and contesting administrative decisions. As it is possible to extract training data, including personal data, from existing pre-trained LLMs, researchers must ensure compliance of their research with data protection law.

The generation of complete legal texts with legal reasoning is not only relevant in the context of administrative decision-making, but for diverse use cases in the legal domain. The prospect of LLMs becoming sufficiently advanced to draft laws, administrative decisions, court rulings,[138] legal submissions, and scholarly essays raises the need for a societal debate on automation. This dialogue should also consider dystopian scenarios where communication in legislative procedures, administrations, courtrooms, and academic legal circles is dominated by LLMs interacting with each other, with minimal human input.

---

[138] On the automation of judicative decisions see Morison and McInerney (n 19) 11.

**Cite this article:** PJ Pesch (2025). "Potentials and Challenges of Large Language Models (LLMs) in the Context of Administrative Decision-Making". *European Journal of Risk Regulation* **16**, 76–95. https://doi.org/10.1017/err.2024.99