Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity

Working Paper (this version: 14 January 2024)

Claudio Novelli[1], Federico Casolari[1], Philipp Hacker[2], Giorgio Spedicato[1], Luciano Floridi[1, 3]

[1] Department of Legal Studies, University of Bologna, Via Zamboni, 27/29, 40126, Bologna, IT
[2] European New School of Digital Studies, European University Viadrina, Große Scharrnstraße 59, 15230 Frankfurt (Oder), Germany
[3] Digital Ethics Center, Yale University, 85 Trumbull Street, New Haven, CT 06511, US

*Email of correspondence author: claudio.novelli@unibo.it

Abstract:

The advent of Generative AI, particularly through Large Language Models (LLMs) like ChatGPT and its successors, marks a paradigm shift in the AI landscape. Advanced LLMs exhibit multimodality, handling diverse data formats, thereby broadening their application scope. However, the complexity and emergent autonomy of these models introduce challenges in predictability and legal compliance. This paper delves into the legal and regulatory implications of Generative AI and LLMs in the European Union context, analyzing aspects of liability, privacy, intellectual property, and cybersecurity. It critically examines the adequacy of the existing and proposed EU legislation, including the Artificial Intelligence Act (AIA) draft, in addressing the challenges posed by Generative AI in general and LLMs in particular. The paper identifies potential gaps and shortcomings in the legislative framework and proposes recommendations to ensure the safe and compliant deployment of generative models, ensuring they align with the EU's evolving digital landscape and legal standards.

1. Overview

Since the release of ChatGPT at the end of 2022, Generative AI in general, and Large Language Models (LLMs) in particular, have taken the world by storm. On a technical level, they can be distinguished from more traditional AI models in various ways.[1] They are trained on vast amounts of text and generate language as output, as opposed to scores or labels in traditional regression or classification (Foster 2022, 4-7; Hacker, Engel, and Mauer 2023). Often, LLMs are marked by their wider scope and greater autonomy in extracting patterns within large datasets. In particular, LLMs' capability for smooth general scalability enables them to generate content by processing a varying range of input from several domains. Many LLMs are multimodal (also called Large

[1] While Generative AI encompasses a wider range of systems than LLMs, their overlapping legal concerns necessitate considering them together. However, we will maintain a focus on LLMs.

Multimodal Models, LMMs), meaning they can process and produce various types of data formats simultaneously: e.g., GPT-4 can handle both text, image, and audio inputs concurrently for generating text.[2] However, while advanced LLMs generally perform well across a broad spectrum of tasks, this comes with highly unpredictable outputs, even for their creators, raising concerns over the lawfulness and accuracy of LLM-generated texts (Ganguli et al. 2022).

The accelerated growth of LLMs, including GPT-4 and Bard, necessitates evaluating the efficacy of existing and forthcoming EU legislation. In this paper, we shall discuss some key legal and regulatory concerns brought up by Generative AI and LLMs regarding liability, privacy, intellectual property, and cybersecurity. The EU's response to these concerns should be contextualized within the guidelines of the Artificial Intelligence Act (AIA) draft, which comprehensively addresses the design, development, and deployment of AI models, including Generative AI within its scope. Where we perceive gaps or flaws in the EU legislation, we will put forth some recommendations to guarantee the safe and lawful use of LLMs.

2. Liability and AI Act

33% of firms view "liability for damage" as the top external obstacle to AI adoption, especially for LLMs, only rivalled by the "need for new laws", expressed by 29% of companies.[3] A new, efficient liability regime may address these concerns by securing compensation to victims and minimizing the cost of preventive measures. In this context, two recent EU regulatory proposals on AI liability may affect LLMs: one updating the existing Product Liability Directive (PLD) for defective products, the other introducing procedures for fault-based liability for AI-related damages through the Artificial Intelligence Liability Directive (AILD).[4] At the moment, the AILD is parked in the legislative process, however.

The two proposals offer benefits for regulating AI liability, including Generative AI and LLMs. First, the scope of the PLD is extended to include all AI systems and AI-enabled goods, except for open-source software, to avoid burdening research and innovation (Rec.13 PLD). This is advantageous as the PLD is the only harmonized European liability law, with a strict liability regime applicable in specific instances. Second, the PLD acknowledges that an AI system can become defective based on knowledge acquired/learned post-deployment, thereby extending liability to such occurrences (Article 6(c) PLD). Third, the AILD addresses claims against non-professional users of AI systems and recognizes violations of fundamental rights among eligible damages. Finally, and perhaps most importantly, both proposals

[2] Moreover, GPT-4 can also generate images thanks to the integration of DALL-E, the AI-powered image generation tool developed by OpenAI.

[3] European Commission, Directorate-General for Communications Networks, Content, and Technology, European enterprise survey on the use of technologies based on artificial intelligence: final report, Publications Office, 2020. The survey refers to the broader category of natural language processing models, pp. 71-72.

[4] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (COM/2022/496 final).

acknowledge AI's opacity and the information imbalance between developers and users or consumers. They introduce disclosure mechanisms and rebuttable presumptions, shifting the burden of proof to developers (AILD Articles 3 and 4; PLD Articles 8 and 9). For instance, under Article 8 PLD and Article 3 AILD, claimants only need to provide plausible evidence of potential harm, while defendants must disclose all relevant information to avoid liability, with non-compliance to this disclosure leading to a (rebuttable) presumption that the defendant has breached its duty of care.

However, both AILD and PLD reveal three major weaknesses when used in the context of Generative AI, largely stemming from their dependence on the AI Act (AIA), which appears ill-suited to govern LLMs effectively. Although the final text of the AIA has not been approved yet, a political agreement reached on December 8th during the EU trilogue provides a fairly stable understanding of its regulatory framework. Despite this stable agreement, it is important to consider additional improvements in the next legislative phases, including the implementing acts, before the AIA is enforced, which is expected to happen no earlier than 2026 (albeit obligations for LLMs may become applicable earlier).

1) *Scope*. The disclosure mechanism and rebuttable presumption in the AILD only apply to high-risk AI systems under the AIA (Art. 6 and Annexes II, III). Hence, the primary issue here is to establish whether, and under what conditions, Generative AI and LLMs might be classified as high-risk systems under the AIA. The EU Parliament version of the AIA, developed on June 14, 2023, no longer automatically designates the category of General-Purpose AI (GPAI) — which includes LLMs — as high risk, as was the case in previous versions. Instead, the EU Parliament version of the risk classification hinges on their downstream application: if used in a high-risk context such as employment or judicial settings (AIA, Annex III), LLMs' deployers are obliged to comply with high-risk system obligations.[5]

Similarly, the trilogue political agreement advocates for differentiating Generative AI requirements based on their downstream application, albeit with a distinct GPAI classification. It establishes a tiered system primarily distinguishing between standard GPAI models and GPAI models with systemic risks. To determine if a model has systemic risks, the amount of floating-point operations (FLOPS) consumed during its training is considered a key factor. LLMs trained using more than 10^25 floating-point operations (FLOPS) would likely be classified as AI models posing systemic risks.[6] The use of FLOPs as an indicator is based on the assumption that greater computational power leads to more complex models with wider societal implications.

Accordingly, the AI Act would impose requirements on "standard" GPAI models, such as providing technical documentation, sharing information with other AI system

[5] An implementing act adopted by the European Commission should clarify how requirements for high-risk AI systems will be applied to general-purpose AI systems.

[6] This is what emerges from the European Commission's Q&A on the Artificial Intelligence Act, updated on December 12 and 14, 2023.

providers, disclosing training data summaries, and complying with copyright laws. Additionally, transparency requirements regarding artificially generated or manipulated content are likely to be incorporated, albeit as part of broader obligations (Clifford Chance 2023).

For GPAI models posing systemic risks, the AI Act would introduce stricter rules, such as model evaluation, systemic risk assessment and mitigation, adversarial testing, serious incident management and corrective measures, robust cybersecurity measures, and potentially energy consumption reporting, as part of an emphasis on AI's environmental sustainability (Clifford Chance 2023). To help GPAI providers follow the AI Act, the Commission also encourages them to work with experts on a code of conduct. Once approved, these codes will help them show they are compliant. This is especially important for outlining how to assess and manage risks for GPAI models with systemic risks. As a result, these LLMs with systemic risks are likely to be subjected to the disclosure mechanism and rebuttable presumption in the AILD.

While these revisions to the initial draft of the AIA represent a positive step toward more effective risk assessment, concerns remain. So, for instance, extending the three-tier classification system to GPAIs–thus mirroring the risk classification based on broad upstream applications as for the other AI systems (e.g., Annex III, AIA)–may fail to account for the peculiarities of downstream applications, potentially leading to over-inclusive or under-inclusive risk categories (Novelli et al. 2023a). Also, the trilogue's two-tier classification of standard and systemic risks for LLMs may be complex, particularly in its definition of systemic risk, which is primarily based on the computational resources used for training, measured in FLOPs. While other aspects may be considered by the Commission, too, FLOPs will be crucial in practice. The reality is that the risks associated with AIs, including LLMs, are multidimensional. They depend on various factors such as the context of application, model architecture, and the quality of training, rather than just the quantity of computational resources used. FLOPs offer only a partial perspective on dangerousness and do not account for how different, non-computational risk factors might interact and potentially lead to cascading failures, including interactions among various LLMs. Finally, the very threshold of 10^25 FLOPs as a risk parameter is questionable (The Future Society 2023) (Moës and Ryan 2023). LLMs with 10^24 or 10^23 FLOPs can be equally risky (e.g., GPT-3; Bard). This is further compounded by the trend towards downsizing LLMs while maintaining high performance and associated risks, such as in the case of Mistral's Mixtral 8x7B model (Hacker 2023b). Again, while this is an ancillary issue as the AI Office will have the power to adjust this parameter, relying solely on FLOPs as a risk indicator remains inadequate.

A second issue, related to scope, involves the alignment between the GPAI and the so-called 'foundation models'. Indeed, the position adopted by the European Parliament in June 2023 has further modified the scope of the AIA.[7] In particular, the

[7] Report 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) and amending certain Union Legislative Acts, doc. A9-0188/2023.

EP version has included foundation models in the regulatory framework at stake. According to the amended text of the AIA, a foundation model is an 'AI system model that is trained on broad data at scale, is designed for the generality of output, and can be adapted to a wide range of distinctive tasks'.[8] While that notion, echoing the definition given in a paper elaborated by a group of scholars at Stanford University (Bommasani et al. 2022), might include LLMs, the proposed resolution recognizes that 'general-purpose AI systems can be an implementation of a foundation model'.[9] As a result, the new notion has not entirely replaced that of GPAIs, which is still present in the proposed text. Foundation models are subject to specific requirements, defining the legal situation of providers and dealing with the assessment and mitigation of possible risks and harms. Importantly, the proposed resolution makes it clear that '[t]hese specific requirements and obligations do not amount to considering foundation model as high-risk AI system…'.[10]

Regarding the first issue of scope, a more use-case-specific and refined approach is necessary to determine the risk level of Generative AI and LLMs. The latest version of the AI Act has taken the first steps toward this goal by introducing the extra layer in Article 6(2), according to which AI systems used in Annex III areas are not automatically qualified as high-risk, but only if they pose a *significant* risk of harm to natural persons or the environment. While the definition of "significant" still lacks granularity, it marks a welcome step toward nuanced risk assessment. Similarly, Article 7 grants the Commission powers to remove or add specific use cases or categories to the high-risk field. A similar move should be made for Generative AI. This entails eschewing the two-tier system solely/primarily anchored in computation potency quantified in FLOPs. Indeed, as the legislative process advances, the focus should shift towards crafting implementing measures under the AIA that incentivize scenario-based risk classification and management (Novelli et al. 2023b; 2023a). This method necessitates a thorough analysis of the LLMs in their specific deployment contexts, considering the potential risks to various assets and individuals. The inherent risk of LLMs may vary significantly based on their end uses, which can introduce additional, external risks. These risks necessitate distinct responsibilities for each entity in the value chain (Hacker, Engel, and Mauer 2023). Moreover, the identification of the stakeholders who may be harmed by LLMs is crucial to this scenario-based risk assessment (Bender et al. 2021).[11] Given the wide array of functions and applications that a single LLM can have within a general field, it is crucial to develop standards that are based on real-world scenarios. Thus, for instance, within the employment sector — a domain classified as high-risk under the AIA — there could be a considerable variation in potential risks between using an LLM for optimizing resume screenings or

[8] Ibid., Article 3, para. 1, point 1c.

[9] Ibid., Rec. 60e.

[10] Ibid, Rec. 60g.

[11] The PLD, which is not tied to the risk categories of the AIA in terms of applicability, cannot do all the work because its provisions apply only to professionals – economic operators – and not to non-professional users like the AILD.

deploying it for automated virtual interviews, where biases might be more prevalent and human supervision might be less effective.

As a result, legislators should use these criteria to determine which LLMs present sufficient risk to fall under the scope of the high-risk provisions of the AI Act and, as a consequence, the AILD. This implies allocating the burden of proof and outlining the obligations of LLM deployers, such as the requirement to conduct a Fundamental Rights Impact Assessment (FRIA).

The second issue of scope should be resolved by overcoming the existing distinction between General Purpose Artificial Intelligence (GPAI) and foundation models. The concept of foundation models renders the GPAI category redundant, at least in a normative framework of the AIA. It should be noted that foundation models, which are developed based on broad-spectrum data and are designed to be adaptable across a wide range of tasks, sometimes inherently encompass functionalities of GPAI, including the attributes seen in LLMs and other types of Generative AI. At other times, LLMs are just built on top of foundational models, targeted at fulfilling very specific and limited purposes (examples could include custom chatbots or task-specific automation tools).

The AIA shows intermittent awareness of this overlap, as seen in provisions like Recommendation 60g (EP Position text). However, the frequent interchangeable use of the terms foundation models and GPAI gives rise to considerable confusion. To mitigate this, a redefined classification is recommended. This classification should use an independent normative label for AI technologies that are developed with large-scale data training and exhibit a wide scope in terms of output generation and adaptability to diverse tasks. Conversely, LLMs and other forms of Generative AI, when used for narrower, more circumscribed purposes, should be treated as the other AI systems listed in the AIA. This categorization should still be tailored to their specific downstream applications, aligning with the modifications previously proposed in this paper (and also with the amendments to the original draft proposed in the compromise text). In essence, the use of the label "foundation models" (or something else) might include Generative AI that showcases a broad function and purpose. However, when their applications are more confined, there is no need to include them under the broader category of GPAI, particularly in the updated normative framework of the AIA. The label "GPAI" might be unnecessary in the AIA context altogether. This approach not only simplifies regulatory classifications but also enhances the effectiveness of governance mechanisms about LLMs.

2) *Defectiveness and fault.* The two directive proposals assume that liability may arise from two different sources–defectiveness (PLD) and fault (AILD)–that are both evaluated by compliance with the requirements of the AIA. Both presume fault/a defect in case of non-compliance with the (high-risk systems) requirements of the AIA (Article 9(2)(b) PLD; Article 4(2) AILD), requirements which could also be introduced at a

later stage by sectoral EU legal instruments.[12] However, these requirements may not be easily met during the development of LLMs: e.g., their lack of a single or specific purpose before adaptation (Bommasani et al. 2022) could hamper the predictions of their concrete impact on the health, safety, and fundamental rights of persons in the Union which are required by the AIA risk management system and transparency obligations (Articles 9 and 13). Moreover, as just mentioned, further requirements are likely to be introduced in the EU regulatory framework concerning foundation models.

To enhance the effectiveness and reliability of LLMs, a necessary recommendation is to combine the conventional AI fault and defectiveness criteria with new methods specifically designed to align with their technical nuances. This may imply that, for LLMs, the compliance requirements for evaluating faults and defectiveness should prioritize techniques for steering the randomness of their non-deterministic outputs over their intended purposes. Indeed, their capability for smooth general scalability enables them to generate content by processing diverse inputs from arbitrary domains with minimal training (Ganguli et al. 2022). To this scope, several techniques might be incentivised by the regulator, also concurrently: e.g., temperature scaling, top-k sampling, prompt engineering, and adversarial training (Hu et al. 2018). Methods for tempering the randomness may also include the so-called regularization techniques, like the dropout, which involves temporarily disabling a random selection of neurons during each training step of LLMs, fostering the development of more robust and generalized features (Lee, Cho, and Kang 2020). Consequently, it prevents the model from overfitting, ensuring more coherent and less random outputs.

Furthermore, compliance requirements for Generative AI and LLMs should also prioritize monitoring measures. These measures would serve to verify that the models operate as planned and to pinpoint and amend any divergences or unfavourable results. For example, calculating the uncertainty of outputs could be instrumental in recognizing situations where the model might be producing highly random results (Xiao et al. 2022). Such information is vital for end-users to have before utilizing an LLM, representing a metric for evaluating the fault of the designers and deployers (or the defectiveness of the same).

3) *Disclosure of evidence*. Both proposals state that the defendant — in our analysis, the deployers and designers of LLMs — must provide evidence that is both relevant and proportionate to the claimant's presented facts and evidence. Shortcomings here concern the scope and the content of such disclosure. First, the PLD and the AIA are misaligned as the former requires evidence disclosure for all AI systems, whereas the Commission's AIA proposal mandates record-keeping obligations only for high-risk systems (Hacker 2023a). Despite the recent introduction of specific guidelines for foundational models, irrespective of their risk levels, we have noted that the classification of LLMs within this category remains somewhat ambiguous. Second,

[12] The dependence on the AIA is less of an issue for the PLD as it has greater harmonization and extensive case law. However, identifying the appropriate safety requirements (Articles 6 and 7) to assess the defectiveness of LLMs remains a challenge.

both the PLD and the AILD do not indicate what type of information must be disclosed. While this issue can be attributed to their status as proposals, it is this gap that should be promptly addressed. Failing to establish clear guidelines on the necessary disclosures might leave the claimants practically unprotected.

Regarding the issue of scope, the requirement to disclose evidence should not be confined to high-risk systems alone. The PLD could potentially adopt the AILD's approach, which broadens the disclosure requirement to include opaque systems that are not classified as high-risk while exempting high-risk systems that already have ample documentation under the AIA (Article 4(4) and (5) AILD). While the content of disclosure might vary based on the system's risk level, maintaining the obligation to disclose is essential.

This leads us to the second point of discussion: the content of disclosure. It should include a report of the damaging incident, noting the exact time and a brief description of its nature. It might include interaction logs and timestamps between users and the LLMs, demonstrating adherence to relevant standards, possibly verified through third-party audit reports (Falco et al. 2021). Moreover, the disclosure should also mirror the sociotechnical structure of LLMs' liability (Novelli, Taddeo, and Floridi 2023; Theodorou and Dignum 2020) and demonstrate that LLMs' training data are representative and well-documented, e.g., in terms of the motivation behind data selection and transparency about the objectives of data collection (Bender et al. 2021; Jo and Gebru 2020). Also, producers might be obligated to use only documentable datasets of an appropriate size for the capabilities of the organization. LLMs operating on restricted datasets–thanks to their few/zero-shot learning skills (Brown et al. 2020)–may instead need to disclose the auxiliary information used for associating observed and non-observed classes of objects.

To conclude, the process of evidential disclosure presupposes that individuals are informed when they are engaging with these models, and consequently, whether they have been adversely affected in specific ways. However, even though the stipulations outlined in the AIA mandate the notification of users during interactions with foundational models, the methodology for user notification remains ambiguous (Ziosi et al. 2023). This is a key point as the efficacy of the disclosure mechanisms hinges on this prerequisite, wherein to lodge claims, users must possess a reasonable basis to suspect harm and furnish substantial details and corroborating evidence to substantiate a potential damages claim. Since the acquisition of this knowledge can present challenges, it is recommended to encourage LLMs' producers to actively notify occurrences of potential harm. This strategy would not only bolster the claimant's ability to access crucial evidence but would also cultivate a more transparent environment within the operational sphere of LLMs. Such incentives might include initiatives like forming alliances with credible third-party organizations, including auditing agencies, to facilitate a thorough disclosure of information (and evidence) concerning the adverse effects linked with the use of LLMs.

3. Privacy and Data Protection

Privacy and data protection pose critical legal hurdles to the development and deployment of Generative AI, as exemplified by the 2023 Italian data authority's (Garante della Privacy) temporary ban on ChatGPT (Hacker, Engel, and Mauer 2023, Technical Report). On an abstract level, an LLM preserves privacy if it discloses confidential information in appropriate contexts and to authorized individuals only. Privacy and data protection are not binary variables and, therefore, what is the right context or the right recipients of the information is a matter of debate. In the context of LLMs, these debates are further complicated due to the diverse purposes, applications, and environments they operate.[13]

LLMs are exposed to privacy and data protection breaches due to pervasive training on (partially) personal data, the memorization of training data, and inversion attacks (Nicholas Carlini et al. 2021). Memorization of data may occur either through overfitting of abundant parameters to small datasets, which reduces the capacity to generalize to new data, or through the optimizing generalization of long-tailed data distributions (Feldman 2021). When the memorized training data contains personal information, LLMs may leak data and disclose it directly. When training data is not memorized, personal information can still be inferred or reconstructed by malicious actors using model inversion attacks, which reverse-engineer the input data to reveal private information (Fredrikson, Jha, and Ristenpart 2015). Against this, the existing privacy-preserving strategies, such as data sanitization and differential privacy, provide limited privacy protection when applied to LLMs (Brown et al. 2022). This raises the question of whether, and how, personal data may be processed to train LLMs–a particularly thorny question concerning sensitive data. Moreover, users may enter private information through prompts, which may resurface in other instances. Some users, in addition, will be minors, for whom specific data protection rules apply.

In our view, this leads to seven main problems at the intersection of data protection and LLMs: the appropriate legal basis for AI training; the appropriate legal basis for processing prompts; information requirements; model inversion, data leakage, and the right to erasure; automated decision-making; protection of minors; and purpose limitation and data minimization. We will then offer some thoughts on potential ways forward.

1) *Legal basis for AI training on personal data*. First and foremost, every processing operation of personal data–be it storage, transfer, copying, or else–needs a legal basis under Article 6 GDPR. For companies without an establishment in the EU, the GDPR still applies if their services are offered in the EU, for example, which is the case for many major LLM products. The GDPR also covers processing before the actual release of the model, i.e., for training purposes (Oostveen 2016). LLMs are typically

[13] For this discussion, we will concentrate on strategies to prevent LLMs from compromising user privacy and personal data, bypassing what makes a context or a recipient. However, an analysis of these issues is done by (Brown et al. 2022).

trained on broad data at scale, with data sources ranging from proprietary information to everything available on the Internet–including personal data, i.e., data that can be related to an identifiable individual (Bommasani et al. 2021). Using this type of data for AI training purposes, hence, is illegal under the GDPR unless a specific legal basis applies. The same holds for any fine-tuning operations after initial pre-training.

a) Consent and the balancing test

The most prominent legal basis in the GDPR is consent (Article 6(1)(a)). However, for large data sets including personal information from a vast group of people unknown to the developers beforehand, eliciting valid consent from each individual is generally not an option due to prohibitive transaction costs (Mourby, Ó Cathaoir, and Collin 2021). Furthermore, using LLMs with web-scraped datasets and unpredictable applications is difficult to square with informed and specific consent (Bommasani et al. 2022). At the same time, requiring data subjects to be informed about the usage of their personal data may slow down the development of LLMs (Goldstein et al. 2023). Hence, for legal and economic reasons, AI training can typically be based only on the balancing test of Article 6(1)(f) GDPR (Zuiderveen Borgesius et al. 2018; Zarsky 2017), according to which the legitimate interests of the controller (i.e., the developing entity) justify processing unless they are overridden by the rights and freedoms of the data subjects (i.e., the persons whose data are used).[14]

Whether the balancing test provides a legal basis is, unfortunately, a matter of case-by-case analysis (Gil Gonzalez and de Hert 2019; Peloquin et al. 2020; Donnelly and McDonagh 2019). Generally, particularly socially beneficial applications will speak in favour of developers; similarly, control is unlikely to prevail if the use of the data for AI training purposes could reasonably be expected by data subjects, Recital 47. That latter criterion, however, will rarely be fulfilled. By contrast, the nature and scope of processing, the type of data (sensitive or not), the degree of transparency towards and control for data subjects, and other factors may tip the balance in the other direction (Hacker, Engel, and Mauer 2023, Technical Report, 2).

For narrowly tailored AI models based on supervised learning strategies, one may argue that AI training is not particularly helpful as it does not, generally, reveal any new information about the data subjects themselves (Hacker 2021; Zarsky 2017; Bonatti and Kirrane 2019). This argument is particularly strong if the model is not passed along to other entities in state-of-the-art IT security making data breaches less likely.

However, this position is difficult to maintain concerning LLMs (Hacker, Engel, and Mauer 2023, Technical Report, 2): these models are generally used by millions of different actors, and models have been shown to reveal personal data through data leakage as well as model inversion (Nicholas Carlini et al. 2021; Bederman 2010; Lehman et al. 2021; Nicolas Carlini et al. 2023). This poses an even greater challenge in fine-tuning scenarios (Borkar 2023).

[14] Another possibility is the purpose change test (Article 6(4) GDPR), not explored further here for space constraints. Note that Article 9 GDPR, in our view, applies in addition.

b) Sensitive Data

To make matters even more complex, a much larger number of personal data pieces than expected may be particularly protected as sensitive data under Article 9 GDPR, under a new ruling of the CJEU. In Meta v. Bundeskartellamt, the Court decided that information need not directly refer to protected categories–such as ethnic or racial origin, religion, age, or health–to fall under Article 9. Rather, it suffices "that data processing allows information falling within one of those categories to be revealed".[15] That case was decided concerning Meta, the parent company of Facebook, based on its vast collection of data tracking users and linking that data with the user's Facebook account.

Arguably, however, as is generally the case in technology-neutral data protection law, the exact method of tracking or identification is irrelevant; the Court held that it does not matter, for example, whether the profiled person is a Facebook user or not.[16] Rather, from the perspective of data protection law, what is decisive is the controller's ability to infer sensitive traits based on the available data–irrespective of whether the operator intends to make that inference. This broader understanding casts a wide net for the applicability of Article 9 GDPR, as machine learning techniques increasingly allow for the deduction of protected categories from otherwise innocuous data points.

Hence, in many cases concerning big data formats, the hypothetical possibility to infer sensitive data potentially brings the processing, for example, for AI training purposes, under the ambit of Article 9. Developers then need to avail themselves of the specific exception in Article 9(2) GDPR. Outside of explicit consent, such an exception will, however, often not be available: Article 9(2) does not contain a general balancing test, in contrast to Article 6(1) GDPR (and the secondary use clause in Article 6(4). The research exemption in Article 9(2)(j) GDPR, for example, is limited to building models for research purposes, but cannot be used to exploit them commercially (cf. Recitals 159 and 162).

Overall, this discussion points to the urgent need to design a novel exemption to Article 9, accompanied by strong safeguards, similar to the one contemplated in Article 10(5) AIA, to balance the societal interest in socially beneficial AI training and development with the protection of individual rights and freedoms, particularly in crucial areas such as medicine, education, or employment. While the TDM exception provides for a specific framework for the use of copyrighted material for AI training purposes, such rules are, unfortunately, entirely lacking under the GDPR.

2) *Legal basis for prompts containing personal data.* The situation is different for prompts containing personal data entered into a trained model. Here, we have to fundamentally distinguish two situations. First, users may include personal information about themselves in prompts, for example, when they ask an LLM to draft an email concerning a specific event, appointment, or task. In this case, consent may indeed work as a legal basis as users have to individually register for the LLM product. During

[15] CJEU, C-252/21, Meta vs. Bundeskartellamt, ECLI:EU:C:2023:537, para. 73.

[16] CJEU, C-252/21, Meta vs. Bundeskartellamt, ECLI:EU:C:2023:537, para. 73.

that procedure, controllers may consent (respecting the conditions for valid consent under Articles 4(11) and 7 GDPR, of course).

The second scenario concerns prompts containing personal information about third parties, i.e., not the person entering the prompt. This situation is more common among users who might not be fully aware of privacy and data protection laws. They might inadvertently include the personal details of others if the task at hand involves these third parties, and they expect the language model to provide personalized responses. Users cannot, however, validly consent for another person (unless they have been explicitly mandated by that person to do just that, which is unlikely). Hence, a similar problem resurfaces as in the AI training or fine-tuning scenario, with the additional twist that the information is provided, and processing initiated, by the user, not the developers. While the user may be regarded as the sole controller, or joint controller together with the company operating the LLM (Article 4(7) GDPR), for the initial storage and transfer of the prompt (i.e., writing and sending the prompt), any further memorization or data leakage is under the sole control of the entity operating the LLM. Hence, under the *Fashion ID* judgment of the CJEU,[17] that operational entity will likely be considered the sole controller, and hence the responsible party (Art. 5(2) GDPR), for any storage, transfer, leakage, or other processing of the third-party-related personal data included in the prompt that occurs after the initial prompting by the user. Again, as in the training scenario, both the third-party-related prompt itself and any additional leakage or storage are not easy to justify under Article 6(1)(f) and, if applicable, Article 9 GDPR.

3) *Information requirements*. Major roadblocks for GDPR-compliant LLMs are Articles 12-15 GDPR, which detail the obligations regarding the information that must be provided to data subjects. These articles pose a unique challenge for LLMs due to the nature and scope of data they process (Hacker, Engel, and Mauer 2023, Technical Report, 2-3).

When considering data harvested from the internet for training purposes, the applicability of Article 14 of the GDPR is crucial. This article addresses the need for transparency in instances where personal data is not directly collected from the individuals concerned. However, the feasibility of individually informing those whose data form part of the training set is often impractical due to the extensive effort required, potentially exempting it under Article 14(5)(b) of the GDPR. Factors such as the volume of data subjects, the data's age, and implemented safeguards are significant in this assessment, as noted in Recital 62 of the GDPR. The Article 29 Working Party particularly notes the impracticality when data is aggregated from numerous individuals, especially when contact details are unavailable (Article 29 Data Protection Working Party 2018, para. 63, example).

Conversely, the processing of personal data submitted by users on themselves in a chat interface (prompts) is not subject to such exemptions. Article 13 of the GDPR explicitly requires that data subjects be informed of several key aspects, including

[17] CJEU, C-40/17, Fashion ID, ECLI:EU:C:2019:629.

processing purposes, the legal basis for processing, and any legitimate interests pursued. Current practices may not have fully addressed these requirements, marking a significant gap in GDPR compliance.

Importantly, the balance between the practical challenges of compliance and the rights of data subjects is delicate. While the concept of disproportionate effort under Article 14(5) GDPR presents a potential exemption, it remains a contentious point, particularly for training data scraping and processing for commercial purposes. In this regard, the data controller, as defined in Article 4(7) of the GDPR, should meticulously document the considerations made under this provision. This documentation is a crucial aspect of the accountability principle enshrined in Article 5(2) of the GDPR. Furthermore, in our view, documents regarding the methods of collecting training data should be made publicly accessible, reinforcing the commitment to GDPR principles.

4) Model inversion, data leakage, and the right to erasure. GDPR compliance for LLMs gets even trickier with concerns about reconstructing training data from the model (model inversion) and unintentional data leaks, especially in light of the right to be forgotten (or right to erasure) under Article 17. Things get further complicated when some argue that LLMs themselves might be considered personal data due to their vulnerability to these attacks (Veale, Binns, and Edwards 2018). Inversion attacks refer to techniques whereby, through specific attacks, individuals' data used in the training of these models can be extracted or inferred. Similarly, the memorization problem, which causes LLMs to potentially output personal data contained in the training data, may be invoked to qualify LLMs themselves as personal data.

The ramifications of classifying the model as personal data are profound and far-reaching. If an LLM is indeed deemed personal data, it implies that data subjects could, in theory, invoke their right to erasure under Article 17 of the GDPR. This right, also known as the 'right to be forgotten' allows individuals to request the deletion of their personal data under specific conditions. In the context of LLMs, this could lead to unprecedented demands for the deletion of the model itself, should it be established that the model contains or constitutes personal data of the individuals.

Such a scenario poses significant challenges for the field of AI and machine learning. The practicality of complying with a request for erasure in this context is fraught with technical and legal complexities. Deleting a model, particularly one that has been widely distributed or deployed, could be technologically challenging and may have significant implications for the utility and functionality of the system. Furthermore, this approach raises questions about the balance between individual rights and the broader benefits of AI technologies. The deletion of entire models, with a potential subsequent economic need to retrain the entire model, also raises intricate questions concerning environmental sustainability given the enormous energy and water consumption of (re-)training LLMs (Hacker 2024).

Although LLM producers, such as OpenAI, claim to comply with the right to erasure, it is unclear how they can do so because personal information may be conveyed in multiple forms in an LLM, which escalates the complexity of identifying and isolating specific data points, particularly when the data is not presented in a

structured format (e.g., phone numbers). Additionally, the removal requests initiated by a single data subject may prove to be inadequate, especially in scenarios where identical information has been circulated by multiple users during their engagements with the LLM (Brown et al. 2022). In other words, the deletion of data from a training dataset represents a superficial solution, as it does not necessarily obliterate the potential for data retrieval or the extraction of associated information encapsulated within the mode's parameters. Data incorporated during the training phase can permeate the outputs generated by certain machine learning models, creating a scenario where original training data, or information linked to the purged data, can be inferred or "leaked" thereby undermining the integrity of the deletion process and perpetuating potential privacy violations (De Cristofaro 2020). At a minimum, this points to the need for more robust and comprehensive strategies to address data privacy within the operational area of LLMs.

5) Automated decision-making. Furthermore, given new CJEU jurisprudence, the use of LLMs might be qualified as automated decision-making processes, a topic scrutinized under Article 22 of the GDPR. This article generally prohibits automated individual decision-making, including profiling, which produces legal effects concerning an individual or similarly significantly affects them, unless specific exceptions apply. These exceptions include explicit consent or the necessity of the decision-making for the entry into, or performance of, a contract.

In cases where LLMs are used for evaluation, such as in recruitment or credit scoring, the importance of this regulation becomes even more significant. A pertinent illustration is provided by the recent ruling in the SCHUFA case by the CJEU.[18] The Court determined that the automated generation of a probability value regarding an individual's future ability to payment commitments by a credit information agency constitutes 'automated individual decision-making" as defined in Article 22. According to the Court, this presupposes, however, that this probability value significantly influences a third party's decision to enter into, execute, or terminate a contractual relationship with that individual.

Extrapolating from this ruling, the automated evaluation or ranking of individuals by LLMs will constitute automated decision-making if it is of paramount importance for the decision at hand–even if a human signs off on it afterward. The legal implications of this are profound. Exemptions from the general prohibition of such automated decision-making are limited to scenarios where there is a specific law allowing the process, explicit consent, or where the automated processing is necessary for contractual purposes, as per Article 22(2) of the GDPR.

Obtaining valid consent in these contexts is challenging, especially considering the power imbalances often present between entities like employers or credit agencies and individuals seeking jobs or credit (Recital 43 GDPR). Therefore, the legality of using LLMs in such situations may largely depend on whether their use can be justified as necessary for the specific task at hand. Arguments based solely on efficiency are

[18] CJEU, C-634/21, QG vs. SCHUFA, ECLI:EU:C:2023:957, para. 73.

unlikely to be sufficient. Instead, those deploying LLMs for such purposes might need to demonstrate tangible benefits to the applicants, such as more reliable, less biased, or more transparent evaluation processes. Absent such a qualification, only specific union or Member State laws, containing sufficient safeguards, may permit such automated decision making (Article 22(2)(b) GDPR).

This has, again, profound implications. The only exemptions from the prohibition available are consent or the necessity of automated decision-making for entering into a contract (Article 22(2) GDPR). Consent will be difficult to obtain validly given the power differential in many situations, particularly between job/credit applicants and companies. Hence, the legality of using LLMs in these scenarios might hinge on an analysis of the necessity of that use for this specific task. Efficiency considerations alone will likely not suffice; rather, LLM users/deployers will likely have to show that applicants benefit from the intervention as well, for example through more reliable, less biased, or more transparent evaluations and rankings.

6) Protection of minors. Finally, the deployment of LLMs has raised significant concerns regarding age-appropriate content, especially given the potential for generating outputs that may not be suitable for minors. Under Article 8(2) GDPR, the controller must undertake "reasonable efforts to verify […] that [children's] consent is given or authorized by the holder of parental responsibility over the child, taking into consideration available technology."

A notable instance of regulatory intervention in this context is the action taken by the Italian Data Protection Authority (Garante per la Protezione dei Dati Personali—GPDP). On March 30, 2023, the GPDP imposed a temporary restriction on OpenAI's processing of data from Italian users, with a particular emphasis on safeguarding minors.[19] This move underscores the increasing scrutiny by data protection authorities on the implications of LLMs in the context of protecting vulnerable groups, especially children (Malgieri 2023).

In response to these concerns, OpenAI, for example, has implemented measures aimed at enhancing the protection of minors. These include the establishment of an age gap and the integration of age verification tools. The effectiveness and robustness of these tools, however, remain an area of keen interest and ongoing evaluation, especially in the rapidly evolving landscape of AI and data protection.

7) *Purpose limitation and data minimization*. Data controllers should collect personal data only as relevant and necessary for a specific purpose (Article 5(b)-(c)). The AIA reflects this, requiring an assessment of data quantity and suitability (Article 10(e)). However, limiting LLMs' undefined range of purposes, which need extensive data for effective training, might be futile and counterproductive.

One approach to address data calibration for open-ended LLM applications is requiring developers to train models on smaller datasets and leverage few/zero-shot

[19] Garante per la Protezione dei Dati Personali, Provvedimento del 30 marzo 2023 [9870832].

learning skills. As an alternative to imposing restrictions on the dataset, however, it could be more beneficial to strengthen privacy-preserving measures proportionally to dataset size. For example, rather than relying solely on pseudo-anonymization and encryption (Article 10 AIA), LLM providers should implement methods like differential privacy to counter adversarial attacks on large datasets (Shi et al. 2022; Plant, Giuffrida, and Gkatzia 2022).

8) Ways forward. To enable LLMs to comply with GDPR data protection standards, we have already suggested a tailored regime under Art. 9(2) GDPR above. Another reasonable step would be to adapt the data governance measures outlined for high-risk systems in the AIA. Some advancement has already been manifested in the revised text proposal for Article 28(b), which delineates the obligations of foundation model providers from the European Parliament perspective: "process and incorporate only datasets that are subject to appropriate data governance measures […] in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation". However, as previously emphasized, not all LLMs qualify as foundation models and, consequently, some LLMs would not be governed by the provisions outlined in Article 28(b), but rather fall under Article 10.

While the revised iteration of the compromise text for Article 10 is extensive, it may also be generic, necessitating the incorporation of more tailored measures or incentives to aptly address the complexities inherent to LLMs (e.g., under Art. 40/41 AIA). These technical standards should be refined by incorporating LLM-specific measures, such as requiring training on publicly available data, wherever possible. A significant portion of these datasets might also take advantage of GDPR's right to be forgotten exceptions for public interest, scientific, and historical research (Article 17(3)(d)). Where these exceptions do not apply, it could be feasible for LLMs to exclusively utilize datasets not contingent upon explicit consent, which are intended for public usage. Hence, the most appropriate way to use these systems could require fine-tuning public data with private information for individual data subjects' local use. This should be allowed to maximize LLMs' potential, as proposed by (Brown et al. 2022).

Other potential strategies to enhance data privacy are: encouraging the proper implementation of the opt-out right by LLM providers and deployers and exploring the potential of machine unlearning (MU) techniques.

Regarding the first strategy, OpenAI has recently made a potentially significant advancement in this direction by releasing a web crawler, named GPTbot, that comes with an opt-out feature for website owners. This feature enables them to deny access to the crawler, as well as customize or filter accessible content, granting them control over the content that the crawler interacts with.[20] This is useful not only for implementing the opt-out right under the EU TDM copyright exception but also under Article 21 GDPR.

[20] However, skepticism about opting-out tools has raised because, for example, individual users opting-out are not the only holder of their sensitive information (Brown et al. 2022).

Turning to the second strategy, MU stands as potentially a more efficient method to fully implement the right to erasure (Nguyen et al. 2022), a critical aspect when dealing with LLMs. Unlike conventional methods that merely remove or filter data from a training set — a process that is often inadequate since the removed data continues to linger in the model's parameters — MU focuses on erasing the specific influence of certain data points on the model, without the need for complete retraining. This technique, therefore, could more effectively enhance both individual and group privacy when using LLMs (Hine et al. 2023; Floridi 2023).

4. Intellectual Property

Contents generated by LLMs result from processing text data such as websites, textbooks, newspapers, scientific articles, and programming codes. Viewed through the lens of intellectual property (IP) law, the use of LLMs raises a variety of theoretical and practical issues[21] that can only be briefly touched upon in this paper, and that the EU legislation seems not yet fully equipped to address. Even the most advanced piece of legislation currently under consideration by the EU institutions–the AIA–does not contain qualified answers to the issues that will be outlined below. The stakes have been raised significantly, however, by several high-profile lawsuits levelled by content creators (e.g., the New York Times; Getty Images) against Generative AI developers, both in the US[22] and in the EU (de la Durantaye 2023).

Within the context of this paper, it is advisable to distinguish between the training of LLMs on the one side and the subsequent generation of outputs on the other. Furthermore, concerning the generation of outputs, it is worthwhile to further differentiate–as suggested, among the others, by the European Parliament[23]–between instances in which LLMs serve as mere instruments to enhance human creativity and situations in which LLMs operate with a significantly higher degree of autonomy. On the contrary, the possibility of protecting LLMs themselves through an intellectual property right will not be discussed in this paper.

1) *Training*. The main copyright issue concerning AI training arises from the possibility that the training datasets may consist of or include text or other materials protected by copyright or related rights (Sartor, Lagioia, and Contissa 2018). Indeed, for text and materials to be lawfully reproduced (or otherwise used within the training process), either the rights holders must give their permission or the law must specifically allow their use in LLM training.

[21] For a general discussion of these issues, see (J.-A. Lee, Hilty, and Liu 2021) and the compendium provided by WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence*, 21 May 2020, WIPO/IP/AI/2/GE/20/1 REV.

[22] See,e.g.,https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/.

[23] Cf. European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), par. 15.

The extensive scale of the datasets used and, consequently, the significant number of right-holders potentially involved render it exceedingly difficult to envision the possibility that those training LLMs could seek (and obtain) an explicit license from all right holders, reproducing the problem of data protection consent. This issue becomes particularly evident when, as often occurs, LLM training is carried out using web scraping techniques, a practice whose legality has been (and continues to be) debated by courts and scholars in Europe (Sammarco 2020; Klawonn 2019), even in terms of potential infringement of the *sui generis* right granted to the maker of a database by Directive 96/9/EC[24]. On the one hand, some content available online, including texts and images, might be subject to permissive licensing conditions–e.g. some Creative Commons licenses–authorizing reproduction and reuse of such content even for commercial purposes. The owner of a website could, on the other hand, include contractual clauses in the Terms and Conditions of the website that prohibit web scraping even when all or some of the website's content is not *per se* protected by intellectual property rights.[25] To mitigate legal risk, LLMs should be suitably capable of autonomously analyzing website Terms and Conditions, thereby discerning between materials whose use has not been expressly reserved by their right-holders and materials that may be freely used (also) for training purposes.

The OpenAI above's GPTbot web crawler which allows website owners to opt-out or filter/customize content access offers a significant technical tool in this context. While it does not eliminate all IP law concerns, it is a proactive measure that could, in the future, set a standard of care that all LLMs' providers might be expected to uphold.[26] Significantly, the foundation model/GPAI rules of the AIA discussed in the trilogue contained precisely an obligation for providers of such systems to establish a compliance system, via technical and organizational measures, capable of recognizing and respecting rightsholder opt-outs (Hacker 2023c). For the moment, it remains unclear, however, if this provision will be contained in the final version of the AI Act. It would be a stop in the right direction, as commercial LLM training without such a compliance system typically amounts to systematic copyright infringement, even under the new and permissive EU law provisions, to which we now turn.

A potential regulatory solution to ensure the lawful use of training datasets would involve applying the text and data mining (TDM) exception provided by Directive 2019/790/EU (DSMD)[27] to the training of LLMs. Indeed, Article 2(2) DSMD defines text and data mining as "any automated analytical technique aimed at analyzing text and data in digital form to generate information which includes but is not limited to

[24] 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 L 77, 27.3.1996, p. 20 – 28.

[25] As clarified by the Court of Justice of the EU in the *Ryanair* case: CJEU, 15 January 2015, case C-30/14 – *Ryanair*, ECLI:EU:C:2015:10.

[26] B. Kinsella "What is GPTBot and Why You Want OpenAI's New Web Crawler to Index Your Content" blogpost in Synthedia available at: https://synthedia.substack.com/p/what-is-gptbot-and-why-you-want-openais?utm_source=profile&utm_medium=reader2.

[27] 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 ("Digital Single Market Directive"), OJ L 130, 17.5.2019, p. 92 – 125.

patterns, trends and correlations". Considering that the training of LLMs certainly encompasses (although it likely extends beyond) automated analysis of textual and data content in digital format to generate information, an argument could be made that such activity falls within the definition provided by the DSM Directive (Dermawan, n.d.). However, the application of the TDM exception in the context of LLMs training raises non-trivial issues (Pesch and Böhme 2023) (Hacker 2021) (see also, more generally, Geiger, Frosio, and Bulayenko 2018; Rosati 2018).

Firstly, where the TDM activity is not carried out by research organizations and cultural heritage institutions for scientific research–e.g., by private companies and/or for commercial purposes–it is permitted under Article 4(3) DSMD only on condition that the use of works and other protected materials "has not been expressly reserved by their right-holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online". This condition underscores our earlier note on the need for LLMs to automatically analyze the Terms and Conditions of websites and online databases.

Secondly, a further element of complexity is that Article 4(2) DSMD stipulates that the reproductions and extractions of content made under Article 4(1) may only be retained "for as long as is necessary for the purposes of text and data mining". In this sense, if one interprets the TDM exception to merely cover the training phase of LLMs (as separate from the validation and testing phases), LLMs should delete copyrighted content used during training immediately after its use. Consequently, these materials could not be employed to validate or test LLMs. In this perspective, to make the text and data mining exception more effective in facilitating LLM development, it is advisable to promote a broad normative interpretation of "text and data mining", encompassing not only the training activity in the strict sense but also the validation and testing of the LLM.

Thirdly, the exception covers only reproductions and extractions, but not modifications of the content–which will often be necessary to bring the material into a format suitable for AI training. Finally, acc. to Art. 7(2) DSMD, the three-step test (Geiger, Griffiths, and Hilty 2008) contained in Art. 5(5) of the InfoSoc Directive 2001/29/EC restricts the scope of the TDM exception. According to this general limit to copyright exceptions, contained as well in international treaties (Oliver 2001; Griffiths 2009), such exceptions apply only "in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder." Importantly, this suggests that the TDM exception cannot justify reproductions that lead to applications that substitute, or otherwise significantly economically compete with, the protected material used for AI training. However, this is, arguably, precisely what many applications are doing (Marcus and Southen 2024). It remains unclear, however, to what extent the three-step-test limits individual applications of the TDM exception in concrete cases before the courts, as opposed to being a general constraint on Member States' competence to curtail the ambit of copyright (Griffiths 2009, 3–4).

As mentioned, legal proceedings have recently been brought in the United States and the EU to contest copyright infringement related to materials used in the training phase

by AI systems[28]. While the outcomes of such cases are not necessarily predictive of how analogous cases might be resolved in the EU–for example, in the US the fair use doctrine could be invoked (Gillotte 2020), which lacks exact equivalents in the legal systems of continental Europe–it will be intriguing to observe the approach taken by courts across the Atlantic. Note, particularly, that these cases may, among other things, be decided by the extent to which AI systems substitute for, i.e., compete with, the materials they were trained on (so-called transformativeness, see, e.g., (Henderson et al. 2023), a consideration that parallels the debate mentioned above in EU law on the interpretation of the three-step-test (and its transposition into Member State law (Griffiths 2009, 3–4)).

2) *Output generation*. It is now worth focusing on the legal issues raised by the generation of outputs by LLMs. In this regard, two different aspects must be primarily addressed: the legal relationship between these outputs and the materials used during the training of LLMs, and the possibility of granting copyright or patent protection to these outputs.

As for the first aspect, it is necessary to assess whether LLM-generated outputs: (a) give rise to the potential infringement of intellectual property rights in the pre-existing materials, (b) qualify as derivative creations based on the pre-existing materials, or (c) can be regarded as autonomous creations, legally independent from the pre-existing materials.

An answer to this complex legal issue could hardly be provided in general and abstract terms, requiring proceeding with a case-by-case assessment, i.e., by comparing a specific LLM-generated output with one or more specific pre-existing materials. Such a comparison could in principle be conducted by applying the legal doctrines currently adopted by courts in cases of copyright or patent infringement (or, when appropriate, the legal doctrines adopted to assess whether a certain work/invention qualifies as a derivative work/invention). In this perspective, indeed, the circumstance that the output is generated by a human creator or an AI system does not make a significant legal distinction.

In general terms, however, the use of protected materials in the training of an LLM does not imply, *per se*, that the LLM-generated outputs infringe upon the intellectual property rights in these materials[29] or qualify as derivative creations thereof. Broadly speaking, an LLM-generated output could infringe upon legal rights in two main ways. First, if the output exhibits substantial and direct similarities to legally protected elements of pre-existing materials, it would likely violate the (reproduction) rights of those materials. Second, if the legally protected aspects or elements of the pre-existing

[28] See, e.g., Z. Small, "Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement", The New York Times, 10 July 2023, available at: https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html; B. Brittain, "Lawsuit says OpenAI violated US authors' copyrights to train AI chatbot", Reuters, 29 June 2023, available at: https://www.reuters.com/legal/lawsuit-says-openai-violated-us-authors-copyrights-train-ai-chatbot-2023-06-29/.

[29] However, some cases might pose more challenges than others: consider, e.g., the case where an AI system is used to create works that involve existing fictional characters (who are *per se* protected).

materials appear in the LLM output through indirect adaptations or modifications, always unauthorized, then this output would likely qualify as a derivative creation from the pre-existing materials (Gervais 2022; Henderson et al. 2023). For instance, the fact that a text generated by an LLM shares the same style as the works of a specific author (as would occur if a prompt such as "write a novel in the style of Dr. Seuss" were used) would not imply, *per se*, an infringement of the intellectual property rights of that author. This is because, in most European legal systems, the literary or artistic style of an author is not an aspect upon which an exclusive right can be claimed.

If, by contrast, an infringement is found in an LLM output, the person prompting the LLM would first and foremost be liable because she directly brings the reproduction into existence. However, the Court of Justice of the European Union (CJEU) has recently determined that if platforms fail to comply with any of three distinct duties of care, they will be directly accountable for violations of the right to publicly communicate a work.[30] These duties amount to i) expeditiously deleting it or blocking access to infringing uploads of which the platform has specific knowledge; ii) putting in place the appropriate technological measures that can be expected from a reasonably diligent operator in its situation to counter credibly and effectively copyright infringements if the platform knows or ought to know, in a general sense, that users of its platform are making protected content available to the public illegally via its platform; iii) not providing tools on its platform specifically intended for the illegal sharing of such content and not knowingly promoting such sharing, including by adopting a financial model that encourages users of its platform illegally to communicate protected content.[31] These duties could–mutatis mutandis–be transposed to LLM developers concerning the right of reproduction (Nordemann 2024, 2023). This would make good sense, both from a normative perspective encouraging active prevention of copyright infringement and from the perspective of the coherence of EU copyright law across technical facilities.[32]

A distinct and further legal issue arises when an LLM-generated output can be regarded as an autonomous creation, legally independent from the pre-existing materials. In this scenario, the question pertains to whether such output may be eligible for protection under IP law, specifically through copyright (in the case of literary, artistic, or scientific works) or through patent protection (in the case of an invention).

As mentioned at the beginning of this paragraph, the fundamental legal problem, here, stems from the anthropocentric stance taken by intellectual property law. While international treaties and EU law do not explicitly state that the author or inventor must be human, various normative hints seem to support this conclusion. In the context of copyright, for instance, for a work to be eligible for protection, it must be

[30] CJEU, Joined Cases C-682/18 and C-683/18, YouTube vs. Cyando, ECLI:EU:C:2021:503.

[31] CJEU, Joined Cases C-682/18 and C-683/18, YouTube vs. Cyando, ECLI:EU:C:2021:503, para. 102. The latter point addresses specifically piracy platforms, not YouTube (para. 96 and 101).

[32] In his case, one would further have to investigate if Art. 17 DSMD constitutes a lex specialis to the more general *Cyando* case (Geiger and Jütte 2021; Leistner 2020).

original, i.e., it must constitute an author's intellectual creation[33]. This requirement is typically interpreted, also by the Court of Justice of the EU, as the work needed to reflect the author's personality (something that AI lacks, at least for now). Patent law takes a less marked anthropocentric approach, but even here, the so-called inventive step–which, together with novelty and industrial applicability, is required for an invention to be patentable–is normatively defined in terms of non-obviousness to a person skilled in the art[34]. The very existence of moral rights (such as the so-called right of paternity) safeguarding the personality of the author or inventor suggests that the subject of protection can only be human.

Considering these succinct considerations, we can return to the initial question, namely whether an LLM-generated output may be eligible for protection under intellectual property law.

The answer to this question is relatively straightforward when the LLM constitutes a mere instrument in the hands of a human creator, or, to put it differently, when the creative outcome is the result of predominantly human intellectual activity, albeit assisted or enhanced by an AI system. In such a scenario, the European Parliament has stressed that where AI is used only as a tool to assist an author in the process of creation, the current IP framework remains fully applicable[35]. Indeed, as far as copyright protection is concerned, the Court of Justice of the EU has made clear in the *Painer* case[36] that it is certainly possible to create copyright-protected works with the aid of a machine or device. A predominant human intellectual activity can be recognized, also based on the CJEU case law, when the human creator using an LLM makes free and creative choices in the phases of conception, execution, and/or redaction of the work (Hugenholtz and Quintais 2021).

A similar conclusion can be drawn regarding the patent protection of inventive outcomes generated with the support of an LLM (Engel 2020). In this perspective, as noted by some scholars, it would likely be necessary to adopt a broader interpretation of the inventive step requirement, which should be understood in terms of non-obviousness to a person skilled in the art assisted by AI, i.e., an AI-aided human expert (Ramalho 2018; Abbott 2018).

An opposite conclusion is often reached when the LLM operates in a substantially autonomous manner. For the sake of clarity, it is necessary to explain the meaning of "autonomous" as used in this context (Dornis 2021). Obviously, in the current state of technology, some degree of human intervention–at the very least in the form of prompts–will always be necessary for an LLM to generate any output. However, the

[33] Cf. Art. 3(1) of the Database Directive; Art. 6 of the Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights ("Term Directive"), OJ L 372, 27.12.2006, p. 12 – 18; Art. 1(3) of the Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs, OJ L 111, 5.5.2009, p. 16 – 22.

[34] Cf. Art. 56 of the European Patent Convention.

[35] Cf. European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), par. 15.

[36] CJEU, 1 December 2011, case C-145/10, *Painer*, ECLI:EU:C:2011:798.

mere formulation of a prompt by a human being is likely insufficient to recognize a substantial human contribution to the creative output generated by the LLM. The fundamental legal aspect is that a notable human contribution must be discernible not in the broader creative process, but specifically in the resulting creative outcome. This condition is not met when human intervention merely involves providing a prompt to an LLM or even when minor modifications, legally insignificant, are made to the creative outcome generated by the LLM (e.g., minor editing of an LLM-generated text). By contrast, a level of IP protection might be appropriate for significant modifications made to the text produced by the LLM.

The conclusion above, which argues against copyright or patent protection for contents generated by LLMs in a substantially autonomous manner, finds confirmation in the positions taken on this issue by, e.g., the US Copyright Office,[37] affirmed by the United States District Court for the District of Columbia,[38] and the European Patent Office[39]. Furthermore, such a conclusion is consistent with the fundamental rationale of intellectual property of promoting and protecting human creativity, as also reflected at the normative level.[40]

However, some authors have observed (sometimes with critical undertones) that a rationale for protecting LLMs autonomously generated content is the need to protect investments, made by individuals and/or organizations, aimed at bringing creative products to the market (Hilty, Hoffmann, and Scheuerer 2021; Geiger, Frosio, and Bulayenko 2018).
In this case, the further issue of determining to whom such intellectual property rights should be granted emerges. Some national legislations–not coincidentally, following the common law tradition, which exhibits a less pronounced anthropocentric character compared to civil law tradition–acknowledge the possibility of protecting computer-created works (Goold 2021)–i.e. works "generated by computer in circumstances such that there is no human author of the work,"[41]–granting the copyright to the person "by whom the arrangements necessary for the creation of the work are undertaken"[42]. The identity of such a person, however, remains somewhat unclear, as this could be,

[37] On 16 March 2023 the US Copyright Office issued formal guidance on the registration of AI-generated works, confirming that "copyright can protect only material that is the product of human creativity": see Federal Register, Vol. 88, No. 51, March 16, 2023, Rules and Regulations, p. 16191.

[38] United States District Court for the District of Columbia [2023]: Thaler v. Perlmutter, No. 22-CV-384-1564-BAH.

[39] On 21 December 2021 the Legal Board of Appeal of the EPO issued a decision in case J 8/20 (DABUS), confirming that under the European Patent Convention (EPC) an inventor designated in a patent application must be a human being.

[40] Cf. recital no. 10 of 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

[41] Cf. Sec. 178 of the UK Copyright, Designs and Patents Act 1988 ("CDP Act").

[42] Cf. Sec. 9(3) of the CDP Act. Similarly, Sec. 11 of the 1997 Copyright Ordinance (Cap. 528) of Hong Kong and Art. 2 of the 1994 New Zealand Copyright Act.

depending on the circumstances, the developer of the LLM, its trainer, or its user, possibly even jointly (Guadamuz 2021).
In civil law systems, while awaiting a potential *ad hoc* regulatory intervention, a possible solution could involve applying to LLM-generated outputs the same principle that applies to works and inventions created by an employee within the scope of an employment contract. In such cases, in most EU legal systems, copyright or patent rights are vested in the employer. Similarly, in situations where the "employee" is artificial, the intellectual property right could be granted to the user of an LLM during entrepreneurial endeavours (Spedicato 2019).

5. Cybersecurity

Although some cybersecurity issues–e.g., privacy breaches–have been already discussed, adversary attacks and misinformation require a specific analysis.

1) *Adversarial attacks*. The complexity and high dimensionality of LLMs make them particularly susceptible to adversarial attacks, i.e., attempts to deceive the model and induce incorrect outputs–such as misclassification–through the feeding of carefully crafted, adversarial data. Cybersecurity is a national competence (Cybersecurity Act, Recital 5) but joint efforts to address it should still be pursued at the EU level, going beyond the general principle of AI robustness. Importantly, the AIA mandates high-risk systems to implement technical measures to 'prevent or control attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws' (Article 15 (4)). The EU's Joint Research Centre has recently unveiled a comprehensive guidance document on cybersecurity measures in the context of AI and LLMs (Joint Research Centre (European Commission) et al. 2023). The European Parliament's draft legislation adds another layer. Article 28b asks foundation model providers to build in "appropriate cybersecurity and safety" safeguards, echoing the two-tiered approach tentatively agreed upon in the trilogue (Hacker 2023c). However, effectively countering adversarial attacks requires careful prioritization and targeting within any AI system, not just high-risk ones.

The AIA's risk levels, based on the likelihood of an AI system compromising fundamental legal values, are not a reliable predictor of vulnerability to adversarial attacks. Some AI deemed as high-risk by the AIA, e.g., for vocational training, may not have those technical traits that trigger adversarial attacks, and vice versa. Therefore, the AIA should provide, through supplementary implementation acts, technical safeguards that are proportionate to the attack-triggers of a specific LLM, independently of the AIA risk levels. Attack-triggers include model complexity, overfitting, linear behaviour, gradient-based optimization, and exposure to universal adversarial triggers like input-agnostic sequences of tokens (Wallace et al. 2021). Finally, novel methods to counter adversarial attacks might involve limiting LLM access to trusted users or institutions and restricting the quantity or nature of user queries (Goldstein et al. 2023).

2) *Misinformation*. LLMs can disseminate misinformation, easily, widely, and at a low cost by attributing a high probability to false or misleading claims. This is mainly due to web-scraped training data containing false or non-factual information (e.g., fictional), which lacks truth value when taken out of context. Other times, an opinion reflecting the majority's viewpoint is misrepresented as truth, despite not being verified facts. Misinformation may facilitate fraud, scams, targeted and non-targeted manipulation (e.g., during elections) (AlgorithmWatch AIForensics 2023), and cyber-attacks (Weidinger et al. 2021; Ranade et al. 2021).

A concerning aspect of natural language processing (NLP) in general is the phenomenon of "hallucinations". It refers to the generation of seemingly plausible text that diverges from the input data or contradicts factual information (Ye et al. 2023). These hallucinations arise due to the models' tendency to extrapolate beyond their training data and synthesize information that aligns with their internal patterns, even if it is not supported by evidence. As a result, while NLP models may produce texts that demonstrate coherence, linguistic fluidity, and a semblance of authenticity, their outputs often lack fidelity to the original input and/or are misaligned with empirical truth and verifiable facts (Ji et al. 2023). This can lead to a situation where uncritical reliance on LLMs results in erroneous decisions and a cascade of negative consequences (Zhang et al. 2023), including the spread of misinformation, especially if false outputs are shared without critical evaluation.

There are different kinds of LLMs' hallucinations (Ye et al. 2023) but we cannot discuss them here in detail. In the recent generation of LLMs–e.g., GPT4 and Bard– the 'Question and Answer' kind is particularly frequent. These hallucinations manifest due to the models' tendency to provide answers even when presented with incomplete or irrelevant information (Ye et al. 2023; Adlakha et al. 2023). A recent study found that hallucinations are particularly common when using LLMs on a wide range of legal tasks (Dahl et al. 2024).

EU legislation lacks specific regulations for LLM-generated misinformation. As LLMs become increasingly integrated into online platforms, expanding the Digital Services Act (DSA) to include them, and mandating online platforms to prevent misinformation, seems the most feasible approach. Also, the project to strengthen the EU Code of Practice on Disinformation (2022) can contribute, though its voluntary adherence reduces its overall effectiveness. Tackling LLM-generated misinformation requires updating both the AIA and the DSA. The DSA contains a range of provisions that can be fruitfully applied to LLMs: e.g., Article 22, which introduces "trusted flaggers" to report illegal content to providers and document their notification (Hacker, Engel and Mauer 2023).

However, it is essential to broaden the DSA's scope and the content subject to platform removal duty, which currently covers only illegal content, as LLM-generated misinformation may be completely lawful (Berz, Engel, and Hacker 2023). Being the most technology-focused regulation, the AIA, or its implementing acts, should tackle design and development guidelines to prevent LLMs from spreading misinformation. Normative adjustments should not focus only on the limitation of dataset size but also

explore innovative strategies that accommodate LLMs' data hunger. Some measures might be the same (or similar to those) mentioned for adversarial attacks — restricting LLM usage to trusted users with limited interactions to prevent online misinformation proliferation[43] — while others may include innovative ideas like fingerprinting LLM-generated texts, training models on traceable radioactive data, or enhancing fact sensitivity using reinforcement learning techniques (Goldstein et al. 2023).

Specific solutions to address hallucinations in LLMs are crucial for mitigating the spread of misinformation and should be employed in policy-related applications. Numerous approaches have been proposed in the literature to address this challenge (Tonmoy et al. 2024; Ye et al. 2023). Some of these solutions are broad strategies that focus on optimizing dataset construction, such as implementing a self-curation phase within the instruction construction process. During this phase, the LLM identifies and selects high-quality demonstration examples (candidate pairs of prompts and responses) to fine-tune the underlying model to better follow instructions (Li et al. 2023). Other strategies address the alignment of LLMs with specific downstream applications–which can benefit from supervised fine-tuning (Chung et al. 2022)–as hallucinations often arise from discrepancies between the model's capabilities and the application's requirements (Ye et al. 2023).

Other approaches are narrower and focused on specific techniques, such as prompt engineering, to optimize the output generated by LLMs. This includes incorporating external authoritative knowledge bases (retrieval-augmented generation) (Kang, Ni, and Yao 2023) or introducing innovative coding strategies or faithfulness-based loss functions (Tonmoy et al. 2024).[44]

Another technical solution to mitigate hallucinations in LLMs worth considering is the Multiagent Debate approach, where multiple LLMs engage in an iterative process of proposing, debating, and refining their responses to a given query (Du et al. 2023). The aim is to achieve a consensus answer that is not only more accurate and factually correct but also preserves the richness of multiple perspectives (Ye et al. 2023). This approach draws inspiration from judicial techniques, particularly cross-examination, to foster a more rigorous examination of the LLMs' responses (Cohen et al. 2023).

6. Conclusion

State-of-the-art Generative AI models in general, and LLMs in particular, exhibit high performance across a broad spectrum of tasks, but their unpredictable outputs raise concerns about the lawfulness and accuracy of the generated content. Overall, EU law seems inadequately prepared to cope with such novelties. Policy proposals include

[43] For instance, the draft legislative proposal of the European Parliament requires that the provider of a foundation model shall demonstrate the reduction and mitigation of reasonably foreseeable risks to democracy and the rule of law (Article 28b).

[44] Which basically means establishing a metric to measure faithfulness, that is, the extent to which a model's outputs align with the input data or established truths.

updating current and forthcoming regulations, especially those encompassing AI broadly, as well as the enactment of specific regulations for Generative AI.

7. References

Abbott, Ryan. 2018. 'Everything Is Obvious'. UCLA Law Review. https://www.uclalawreview.org/everything-is-obvious/.

Adlakha, Vaibhav, Parishad Behnam Ghader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. 'Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering'. arXiv. https://doi.org/10.48550/arXiv.2307.16877.

AlgorithmWatch AIForensics. 2023. 'An Analysis of Microsoft's Bing Chat Generative AI and Elections: Are Chatbots a Reliable Source of Information for Voters?'

Article 29 Data Protection Working Party. 2018. "Guidelines on Transparency under Regulation 2016/679, WP260 rev.01."

Bederman, David J. 2010. 'The Souls of International Organizations: Legal Personality and the Lighthouse at Cape Spartel'. In International Legal Personality. Routledge.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜'. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Berz, Amelie, Andreas Engel, and Philipp Hacker. 2023. "Generative KI, Datenschutz, Hassrede und Desinformation–Zur Regulierung von KI-Meinungen." Zeitschrift für Urheber- und Medienrecht: 586.

Biderman, Stella, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. 2023. "Emergent and predictable memorization in large language models." arXiv preprint arXiv:2304.11158.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2022. 'On the Opportunities and Risks of Foundation Models'. arXiv. https://doi.org/10.48550/arXiv.2108.07258.

Bonatti, Piero A., and Sabrina Kirrane. 2019. 'Big Data and Analytics in the Age of the GDPR'. 2019 IEEE International Congress on Big Data (BigDataCongress), July, 7–16. https://doi.org/10.1109/BigDataCongress.2019.00015.

Borkar, Jaydeep. 2023. What Can We Learn from Data Leakage and Unlearning for Law?

Brown, Hannah, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. 'What Does It Mean for a Language Model to Preserve Privacy?' In 2022 ACM Conference on Fairness, Accountability, and Transparency, 2280–92. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3534642.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. 'Language Models Are Few-Shot Learners'. arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al. 2021. 'Extracting Training Data from Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2012.07805.

Carlini, Nicolas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. 'Extracting Training Data from Diffusion Models'. In , 5253–70. https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, et al. 2022. 'Scaling Instruction-Finetuned Language Models'. arXiv. https://doi.org/10.48550/arXiv.2210.11416.

Clifford Chance. 2023. 'The EU's AI Act: What Do We Know about the Critical Political Deal?' https://www.cliffordchance.com/content/cliffordchance/briefings/2023/12/the-eu-s-ai-act--what-do-we-know-about-the-critical-political-de.html.

Cohen, Roi, May Hamri, Mor Geva, and Amir Globerson. 2023. 'LM vs LM: Detecting Factual Errors via Cross Examination'. arXiv. https://doi.org/10.48550/arXiv.2305.13281.

De Cristofaro, Emiliano. 2020. 'An Overview of Privacy in Machine Learning'. arXiv. https://doi.org/10.48550/arXiv.2005.08679.

de la Durantaye, Katharina. 2023. "'Garbage In, Garbage Out'-Die Regulierung generativer KI durch Urheberrecht." ZUM 10 (2023): 645-660.

Dermawan, Artha. n.d. 'Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Nonenjoyment" Purposes?' The Journal of World Intellectual Property n/a (n/a). Accessed 13 August 2023. https://doi.org/10.1111/jwip.12285.

Donnelly, Mary, and Maeve McDonagh. 2019. 'Health Research, Consent, and the GDPR Exemption'. European Journal of Health Law 26 (2): 97–119. https://doi.org/10.1163/15718093-12262427.

Dornis, Tim W. 2021. 'Of "Authorless Works" and "Inventions without Inventor"– The Muddy Waters of "AI Autonomy" in Intellectual Property Doctrine'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3776236.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. 'Improving Factuality and Reasoning in Language Models through Multiagent Debate'. arXiv.Org. 23 May 2023. https://arxiv.org/abs/2305.14325v1.

Durantaye, Katharina de la. 2023. '"Garbage In, Garbage Out" - Die Regulierung Generativer KI Durch Urheberrecht'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4571908.

Engel, Andreas. 2020. 'Can a Patent Be Granted for an AI-Generated Invention?' GRUR International 69 (11): 1123–29. https://doi.org/10.1093/grurint/ikaa117.

Falco, Gregory, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, et al. 2021. 'Governing AI Safety through Independent Audits'. Nature Machine Intelligence 3 (7): 566–71. https://doi.org/10.1038/s42256-021-00370-7.

Feldman, Vitaly. 2021. 'Does Learning Require Memorization? A Short Tale about a Long Tail'. arXiv. https://doi.org/10.48550/arXiv.1906.05271.

Floridi, Luciano. 2023. 'Machine Unlearning: Its Nature, Scope, and Importance for a "Delete Culture"'. Philosophy & Technology 36 (2): 42. https://doi.org/10.1007/s13347-023-00644-5.

Foster, David. 2022. Generative Deep Learning. O'Reilly

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. 2015. 'Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures'. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1322–33. CCS '15. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2810103.2813677.

Ganguli, Deep, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, et al. 2022. 'Predictability and Surprise in Large Generative Models'. In 2022 ACM Conference on Fairness, Accountability, and Transparency, 1747–64. https://doi.org/10.1145/3531146.3533229.

Geiger, Christophe, and Bernd Justin Jütte. 2021. 'Towards a Virtuous Legal Framework for Content Moderation by Digital Platforms in the EU? The Commission's Guidance on Article 17 CDSM Directive in the Light of the YouTube/Cyando Judgement and the AG's Opinion in C-401/19'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3889049.

Geiger, Christophe, Giancarlo Frosio, and Oleksandr Bulayenko. 2018. 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright

in the Digital Single Market - Legal Aspects'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3160586.

Gervais, Daniel. 2022. 'AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines'. Seton Hall Law Review 53 (4). https://scholarship.law.vanderbilt.edu/faculty-publications/1263.

Gil Gonzalez, Elena, and Paul de Hert. 2019. 'Understanding the Legal Provisions That Allow Processing and Profiling of Personal Data—an Analysis of GDPR Provisions and Principles'. ERA Forum 2019 (4): 597–621. https://doi.org/10.1007/s12027-018-0546-z.

Gillotte, Jessica. 2020. 'Copyright Infringement in AI-Generated Artworks'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3657423.

Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations'. arXiv. https://doi.org/10.48550/arXiv.2301.04246.

Goold, Patrick Russell. 2021. 'The Curious Case of Computer-Generated Works under the Copyright, Designs and Patents Act 1988'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4072004.

Griffiths, Jonathan. 2009. 'The "Three-Step Test" in European Copyright Law - Problems and Solutions'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=1476968.

Guadamuz, Andres. 2021. 'Do Androids Dream of Electric Copyright? Comparative Analysis of Originality in Artificial Intelligence Generated Works'. In Artificial Intelligence and Intellectual Property, edited by Jyh-An Lee, Reto Hilty, and Kung-Chung Liu, 0. Oxford University Press. https://doi.org/10.1093/oso/9780198870944.003.0008.

Hacker, Philipp, Andreas Engel, and Marco Mauer. 2023. 'Regulating ChatGPT and Other Large Generative AI Models'. arXiv. https://doi.org/10.48550/arXiv.2302.02337.

Hacker, Philipp. 2021. 'A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act'. Law, Innovation and Technology 13 (2): 257–301. https://doi.org/10.1080/17579961.2021.1977219.

———. 2023a. 'The European AI Liability Directives–Critique of a Half-Hearted Approach and Lessons for the Future'. Computer Law & Security Review 51 (November): 105871. https://doi.org/10.1016/j.clsr.2023.105871.

———. 2023b. 'What's Missing from the EU AI Act: Addressing the Four Key Challenges of Large Language Models'. Verfassungsblog, December. https://doi.org/10.17176/20231214-111133-0.

——— 2023c. 'Statement on the AI Act Trilogue Results', Working Paper, on arxiv.

——— 2024. "Sustainable AI Regulation." Common Market Law Review (forthcoming), https://arxiv.org/abs/2306.00292.

Henderson, Peter, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. 'Foundation Models and Fair Use'. arXiv. https://doi.org/10.48550/arXiv.2303.15715.

Hilty, Reto M, Jörg Hoffmann, and Stefan Scheuerer. 2021. 'Intellectual Property Justification for Artificial Intelligence'. In Artificial Intelligence and Intellectual Property, edited by Jyh-An Lee, Reto Hilty, and Kung-Chung Liu, 0. Oxford University Press. https://doi.org/10.1093/oso/9780198870944.003.0004.

Hine, Emmie, Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2023. 'Supporting Trustworthy AI Through Machine Unlearning'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4643518.

Hu, Zhiting, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2018. 'Toward Controlled Generation of Text'. arXiv. https://doi.org/10.48550/arXiv.1703.00955.

Hugenholtz, P. Bernt, and João Pedro Quintais. 2021. 'Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?' IIC - International Review of Intellectual Property and Competition Law 52 (9): 1190–1216. https://doi.org/10.1007/s40319-021-01115-0.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. 'Survey of Hallucination in Natural Language Generation'. ACM Computing Surveys 55 (12): 248:1-248:38. https://doi.org/10.1145/3571730.

Jo, Eun Seo, and Timnit Gebru. 2020. 'Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning'. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 306–16. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372829.

Joint Research Centre (European Commission), Henrik Junklewitz, Ronan Hamon, Antoine-Alexandre André, Tatjana Evas, Josep Soler Garrido, and Ignacio Sanchez Martin. 2023. Cybersecurity of Artificial Intelligence in the AI Act: Guiding Principles to Address the Cybersecurity Requirement for High Risk AI Systems. LU: Publications Office of the European Union. https://data.europa.eu/doi/10.2760/271009.

Kang, Haoqiang, Juntong Ni, and Huaxiu Yao. 2023. 'Ever: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification'. arXiv. https://doi.org/10.48550/arXiv.2311.09114.

Kinsella, Bret. 2023. 'What Is GPTBot and Why You Want OpenAI's New Web Crawler to Index Your Content'. Substack newsletter. Synthedia (blog). 7 August 2023. https://synthedia.substack.com/p/what-is-gptbot-and-why-you-want-openais?utm_medium=reader2.

Klawonn, Thilo. 2019. 'Urheberrechtliche Grenzen Des Web Scrapings (Web Scraping under German Copyright Law)'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3491192.

Lee, Cheolhyoung, Kyunghyun Cho, and Wanmo Kang. 2020. 'Mixout: Effective Regularization to Finetune Large-Scale Pretrained Language Models'. arXiv. https://doi.org/10.48550/arXiv.1909.11299.

Lee, Jyh-An, Reto Hilty, and Kung-Chung Liu, eds. 2021. 'Artificial Intelligence and Intellectual Property'. In, 0. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198870944.003.0001.

Lehman, Eric, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2021. 'Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?' arXiv. https://doi.org/10.48550/arXiv.2104.07762.

Leistner, Matthias. 2020. 'European Copyright Licensing and Infringement Liability Under Art. 17 DSM-Directive Compared to Secondary Liability of Content Platforms in the U.S.–Can We Make the New European System a Global Opportunity Instead of a Local Challenge?' SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3572040.

Li, Xian, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. 'Self-Alignment with Instruction Backtranslation'. arXiv. https://doi.org/10.48550/arXiv.2308.06259.

Malgieri, Gianclaudio. 2023. Vulnerability and Data Protection Law. Oxford Data Protection & Privacy Law. Oxford, New York: Oxford University Press.

Marcus, Gary, and Reid Southen. 2024. 'Generative AI Has a Visual Plagiarism Problem', 6 January 2024. https://spectrum.ieee.org/midjourney-copyright.

Moës, Nicolas, and Frank Ryan. 2023. Heavy is the Head that Wears the Crown. A risk-based tiered approach to governing General Purpose AI. The Future Society.

Mourby, Miranda, Katharina Ó Cathaoir, and Catherine Bjerre Collin. 2021. 'Transparency of Machine-Learning in Healthcare: The GDPR & European Health Law'. Computer Law & Security Review 43 (November): 105611. https://doi.org/10.1016/j.clsr.2021.105611.

Nguyen, Thanh Tam, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. 'A Survey of Machine Unlearning'. arXiv. https://doi.org/10.48550/arXiv.2209.02299.

Nordemann, Jan Bernd. 2023. "Neu: Täterschaftliche Haftung von Hostprovidern im Urheberrecht bei (Verkehrs-)Pflichtverletzungen im Internet." ZUM: 806-816.

Novelli, Claudio, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2023a. 'How to Evaluate the Risks of Artificial Intelligence: A Proportionality-Based, Risk Model for the AI Act'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4464783.

———. 2023b. 'Taking AI Risks Seriously: A New Assessment Model for the AI Act'. AI & SOCIETY, July. https://doi.org/10.1007/s00146-023-01723-z.

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. 2023. 'Accountability in Artificial Intelligence: What It Is and How It Works'. AI & SOCIETY, February. https://doi.org/10.1007/s00146-023-01635-y.

Oliver, Jo. 2001. "Copyright in the WTO: the panel decision on the three-step test." Colum. JL & Arts 25: 119.

Oostveen, Manon. 2016. 'Identifiability and the Applicability of Data Protection to Big Data'. International Data Privacy Law 6 (4): 299–309. https://doi.org/10.1093/idpl/ipw012.

Peloquin, David, Michael DiMaio, Barbara Bierer, and Mark Barnes. 2020. 'Disruptive and Avoidable: GDPR Challenges to Secondary Research Uses of Data'. European Journal of Human Genetics 28 (6): 697–705. https://doi.org/10.1038/s41431-020-0596-x.

Pesch, Paulina Jo, and Rainer Böhme. 2023. "Artpocalypse now?–Generative KI und die Vervielfältigung von Trainingsbildern." GRUR: 997-1007.

Ramalho, Ana. 2018. 'Patentability of AI-Generated Inventions: Is a Reform of the Patent System Needed?' SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3168703.

Ranade, Priyanka, Aritran Piplai, Sudip Mittal, Anupam Joshi, and Tim Finin. 2021. 'Generating Fake Cyber Threat Intelligence Using Transformer-Based Models'. arXiv. https://doi.org/10.48550/arXiv.2102.04351.

Rosati, Eleonora. 2018. "The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market: technical aspects." European Parliament.

Sammarco, Pieremilio. 2020. 'Creatività Artificiale, Mercato e Proprietà Intellettuale'. Diritto Dell'informazione e Dell'informatica 35 (2). https://cris.unibo.it/handle/11585/716539.

Sartor, Giovanni, Francesca Lagioia, and Giuseppe Contissa. 2018. 'The Use of Copyrighted Works by AI Systems: Art Works in the Data Mill'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3264742.

Spedicato, Giorgio. 2019. 'L'attività Di Web Scraping Nelle Banche Dati Ed Il Riuso Delle Informazioni'. Rivista Di Diritto Industriale 4–5: 253–307.

The Future Society. 2023. EU AI Act Compliance Analysis: General-Purpose AI Models in Focus. Report.

Theodorou, Andreas, and Virginia Dignum. 2020. 'Towards Ethical and Socio-Legal Governance in AI'. Nature Machine Intelligence 2 (1): 10–12. https://doi.org/10.1038/s42256-019-0136-y.

Tonmoy, S. M. Towhidul Islam, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. 'A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2401.01313.

Veale, Michael, Reuben Binns, and Lilian Edwards. 2018. 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law'. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2133): 20180083. https://doi.org/10.1098/rsta.2018.0083.

Wallace, Eric, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2021. 'Universal Adversarial Triggers for Attacking and Analyzing NLP'. arXiv. https://doi.org/10.48550/arXiv.1908.07125.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. 'Ethical and Social Risks of Harm from Language Models'. arXiv. https://doi.org/10.48550/arXiv.2112.04359.

Xiao, Yuxin, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. 'Uncertainty Quantification with Pre-Trained Language Models: A Large-Scale Empirical Analysis'. arXiv. https://doi.org/10.48550/arXiv.2210.04714.

Ye, Hongbin, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. 'Cognitive Mirage: A Review of Hallucinations in Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2309.06794.

Zarsky, Tal. 2017. 'Incompatible: The GDPR in the Age of Big Data'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3022646.

Zhang, Muru, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. 'How Language Model Hallucinations Can Snowball'. arXiv. https://doi.org/10.48550/arXiv.2305.13534.

Ziosi, Marta, Jakob Mökander, Claudio Novelli, Federico Casolari, Mariarosaria Taddeo, and Luciano Floridi. 2023. 'The EU AI Liability Directive: Shifting the Burden From Proof to Evidence'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4470725.

Zuiderveen Borgesius, Frederik, Sanne Kruikemeier, Sophie Boerman, and Natali Helberger. 2018. 'Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the

ePrivacy Regulation'. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3141290.