Can Copyright be Reduced to Privacy?

Niva Elkin-Koren[1], Uri Hacohen[1], Roi Livni[2], and Shay Moran[3]

[1]Faculty of Law, Tel Aviv University
[2]Department of Electrical Engineering, Tel Aviv University
[3]Departments of Mathematics and Computer Science, Technion

May 25, 2023

Abstract

There is an increasing concern that generative AI models may produce outputs that are remarkably similar to the copyrighted input content on which they are trained. This worry has escalated as the quality and complexity of generative models have immensely improved, and the availability of large datasets containing copyrighted material has increased. Researchers are actively exploring strategies to mitigate the risk of producing infringing samples, and a recent line of work suggests to employ techniques such as differential privacy and other forms of algorithmic stability to safeguard copyrighted content.

In this work, we examine the question whether algorithmic stability techniques such as differential privacy are suitable to ensure the responsible use of generative models without inadvertently violating copyright laws. We argue that there are fundamental differences between privacy and copyright that should not be overlooked. In particular we highlight that although algorithmic stability may be perceived as a practical tool to detect copying, it does not necessarily equate to copyright protection. Therefore, if it is adopted as standard for copyright infringement, it may undermine copyright law intended purposes.

1 Introduction

Recent advancements in Machine Learning have sparked a wave of new possibilities and applications that could potentially transform various aspects of our daily lives and revolutionize numerous professions through automation. However, training such algorithms relies heavily on extensive content, either annotated or generated by individuals who may be impacted by these algorithms. Consequently, the identification and determination of when and how content can be used within this framework without infringing upon individuals' legal rights have become a pressing challenge. One area where this issue arises prominently is in the operation of generative models, which take human-produced content—much of it copyrighted– as input and are expected to generate "similar" content. For instance, consider a machine that observes images and is tasked with producing new images that resemble the input. In this context, the fundamental question arises:

When does the content generated by a machine (output content) infringe copyright in the training set (input content).

This question is not purely theoretical, as various aspects of this problem have already become subjects of legal disputes in recent years. In 2022 a class action was filed against Microsoft, GitHub, and OpenAI, claiming that their code-generating systems, Codex and Copilot, infringed copyright in the licensed code that the system was allegedly trained on. Similarly, in another class action, against Stable Diffusion, Midjourney, and DeviantArt, plaintiffs argue that by training their system on web-scraped images, the defendant infringe the rights of millions of artists. Allegedly, the images these systems produce, in response to prompts provided by the systems' users, are based entirely on the training images, which belong to plaintiffs, and, as such, are considered unauthorized derivative works of the plaintiffs' images.

A preliminary question that arises is whether it is lawful to make use of copyrighted content in the course of training [24, 29, 30]. There are compelling arguments to suggest that such intermediary copying might be considered fair use. For example, Google's Book Search Project—entailing the mass digitization of copyrighted books from university library collections to create a searchable database of millions of books—was held by US court to be fair use [2]. 2015).] Then, there is a claim that generative models reproduce protected copyright expressions from the input content on which the model was trained on in response to users' prompts. But to claim that the output content of generative models infringes her copyright, a plaintiff must prove both that the model had access to the protected expression that originated from her work, and that the alleged copy is substantially similar to her work.

To address these challenges, recent studies have proposed measurable metrics to quantify the finding of a copyright infringement [15, 17, 36, 38]. One approach, [15, 38] asserts that copyright is not violated if it is reasonable for the machine to output the content even when no access to the protected content was provided. The argument can be illustrated as follows: Suppose that Alice outputs content A and Bob claims it plagiarizes content B. Alice may argue that she never saw content B, and would reason that this means she did not infringe Bob's copyright. However, since Alice must observe some content, a second line of defense could be that "**had** she never saw B" she would still be likely to produce A.

The above argument was exemplified by Bousquet et al. [15] that interprets differential-privacy in the above manner. Subsequently, Vyas et al. [38] presented a certain generalization, in the form of a *near free access* (NAF) notion that can potentially allow a more versatile notion of copyright protection.

As far as differential privacy, certain traits of copyright law makes it challenging to reduce the problem to a question of privacy. To begin with, an important element of copyright law in the United States is that it has a utilitarian rationale, seeking to promote the creation and deployment of creative works [5, 9]. It is important, then, that any interpretation of copyright, or for that matter any quantifiable measure for copyright, will be aligned with these objectives. In particular, while the law delineates a set of exclusive rights to the creators of original expressions, it must ensure sufficient creative space for current and future creators [35]. As such, certain issues already distinguish copyright law from privacy defined by criteria such as algorithmic stability. Copyright is limited in time, and once protection has expired the work enters the public domain and is free for all to use without authorization [31]. This issue, though, can be modeled by a distinction between private and public data (or protected and non-protected data). More imminent, though, to achieve its goal, copyright law excludes certain subject matter from protection (e.g. ideas, methods of operation, facts), as they are considered raw material for cultural expression. In distinction, requirements such as privacy protect content and not expression, which in turn can be misaligned with the original motivation of the law.

Another distinction from privacy is that copyright law further encourages the use of copyrighted materials by exempting several types of transformative uses, such as quotations, parodies, and some other fair uses such as learning and research [33]. The fair use doctrine serves as a check on copyright, to ensure it does not stifle the very creativity copyright law seeks to foster. Fair use is also considered one of the safety valves which allows copyright protection to coexist with freedom of expression [32].

In this study we initiate a discussion about the challenges involved in providing a rigorous definition that captures the concept of copyright. We commence by formally comparing different proposed notions of copyright (in particular, differential privacy and NAF) and examine their close connection to algorithmic stability. Subsequently, we argue that any approach following this line of reasoning encounters significant obstacles in modeling copyright as understood within the legal context. In more detail, we argue that algorithmic stability strategies fail to account for some key features of copyright law that intend to preserve copyright delicate balance. We identify several major gaps between algorithmic stability strategies and copyright doctrine, demonstrating why applying such strategies may fail to account for essential copyright concepts. Therefore, we argue, that if algorithmic stability techniques are adopted as a standard for copyright infringement, they may undermine copyright law intended goals. We further propose a different approach to using quantified measures in copyright disputes, to better serve and reconcile copyright trade-offs.

1.1 Related Work

A growing number of researchers in recent years explore how to address legal problems by applying theories and methods of computer science. This literature seek to narrow the gap between the vague and abstract concepts used by law and use mathematical models to offer more rigor, coherent and scalable definitions into issues such as privacy [19], or fairness and discrimination [21, 27] In the context of generative models Carlini et al. [17], Haim et al. [25] explore whether generative diffusion models memorize protected works that appeared in the models' training set. This can be considered as a preliminary question to the problem of copyright. However, as discussed, memorization of the input content does not necessarily equates with copyright infringement, and we are thus required to propose measurable metrics and quantified measures for copyright key limiting concepts.

There is also active and thought-provoking discussion on how ML technologies are reshaping our understanding of copyright within the realm of law. Asay [14] explores the question of whether AI system outputs should be subject to copyright. Our focus, though, is on the legitimacy of using copyrighted material by AI. Additionally, as discussed, Grimmelmann [24], Lemley and Casey [30] explore the implications of copyright law for literary machines that extract content and manage databases of information.

The works of Bousquet et al. [15], Vyas et al. [38] which rely on privacy/privacy-like notions, is the main focus of our work. An alternative approach taken by Scheffler et al. [36] proposes a framework to test substantial similarity by comparing Kolmogorov-Levin complexity with and without access to the original copyright work. Beyond algorithmic challenges, to apply a substantial similarity test, though, one has to provide a distinction between protected expressions and non-protected ideas, which may in some cases be the crucial challenge that we might want to solve. Another approach Franceschelli and Musolesi [23] suggests to use generative learning techniques to assess creativity. Henderson et al. [26] seek to develop strategies to be applied to generative models to ensure they satisfy the same fair use standard as in human discretion. The application of this solution may not be possible, though, in cases where little to no open source or fair use data is readily available.

2 Algorithmic stability as a surrogate for copyright

In this section, our focus is to introduce and discuss two notions of algorithmic stability: near-access-freeness (NAF) and differential privacy (DP); these two notions were specifically investigated in the realm of training methods aimed at safeguarding copyrighted data.

Both NAF and DP adhere to a shared form of stability: they ensure that the resulting model, denoted as q, satisfies a safety condition with respect to each copyrighted data instance, denoted as c. This safety condition guarantees the existence of a “safe model”, denoted by q_c, which does not infringe the copyright of data c, and importantly, q exhibits sufficient similarity to q_c. Consequently, both NAF and DP guarantee that p itself does not violate the copyright of the respective data instance c.

Formally, we consider a standard setup of an unknown distribution $\mathcal{D}$, and a generative algorithm A. The algorithm A, gets as an input a training set of i.i.d samples $S = \{z_1, \ldots, z_m\} \in Z^m \sim D^m$, and outputs a model $p_S^A = A(S)$, which is a distribution supported on Z. For simplicity, we will assume here that Z is a discrete finite set, but of arbitrary size. Vyas et al. [38] consider a more general variant in which the output posterior is dependent on a “prompt” x, and A outputs a mapping $p^{(A_S)}(\cdot|x)$ that may be regarded as a mapping from prompts to posteriors. For our purposes there is no loss in generality in assuming that p is “promptless”, and our results easily extend to the promptful case, by thinking of each prompt as inducing a different algorithm when we hard-code the prompt into the algortihm.

Differential Privacy A is said to be (α, β)-differentially private [20] if for every pair of input datasets S, S' that differ on a single datapoint, we have that for every event E:

$$\mathbb{P}(A(S) \in E) \leq e^{\alpha}\mathbb{P}(A(S') \in E) + \beta \text{ and } \mathbb{P}(A(S') \in E) \leq e^{\alpha}\mathbb{P}(A(S) \in E) + \beta \tag{1}$$

The concept of privacy, viewed as a measure of copyright, can be explained as follows: Let’s consider an event, denoted as E, which indicates that the generative model produced by A violates the copyright of a protected content item c. The underlying assumption (which is criticized below) is that if the model has not been trained on c, the occurrence of event E is highly improbable. Thus, we can compare the likelihood of the event E when c is present in the sample S with the likelihood of E when c is not included in a neighboring sample S' (which is otherwise identical to S). If A satisfies the condition stated in equation Eq. (1), then the likelihood of event E remains extremely low, even if c happened to be present once in its training set.

Near Access Freeness There are several shortcomings of the notion of differential privacy that have been identified. Some of these are reiterated in Section 3. Vyas et al. [38] proposed the notion of Near-Access Freeness (NAF) that relaxes differential privacy in several aspects. Formally, NAF (or more accurately NAF w.r.t safe function safe and Δ_{max} is defined as follows: First, we assume a mapping safe that assigns to each protected content c a model q_c which is considered safe in the sense that it does not breach the copyright of c. The function safe, for example, can assign c to a model that was trained on a sample that does not contain c. Several safe functions have been suggested in [38].

A model p is considered α-NAF if the following inequality holds simultaneously for every protected content c and every z:

$$p(z) \leq e^{\alpha} q_c(z). \tag{2}$$

The intuition behind NAF is very similar to the one behind DP, however there are key differences that can, in principle, help it circumvent the stringency of DP.

1. The first difference between NAF and DP is that the NAF framework allows more flexibility by picking the ’safe’ function. Whereas DP is restricted to a safe model corresponding to training the learning algorithm on a neighboring sample excluding the content c.

2. A second difference is the fact that NAF is one sided (see Eq. (2)), in contrast with DP which is symmetric (see Eq. (1)). Note that one-sidedness is indeed more aligned with the requirement of copyright which is non-symmetric.

3. NAF makes the distinction between content-safety and model-safety [38]. In more detail, the NAF notion requires that the output model is stable. This is in contrast with privacy that requires stability of the posterior distribution over the output models. In this sense the notion of NAF is more akin to *prediction differential privacy* [19] then to differential privacy.

4. Finally, NAF poses constraints on the model outputted by the learning algorithm (each constraint corresponds to a prespecified *safe model*). This is in contrast with privacy which does not restrict the output model, but requires stability of the posterior distributions over output models. This distinction may seem minor but it can lead to peculiarities. For example, an algorithm that is completely oblivious to its training set and that always outputs original content can still violate the requirements of NAF. To see this, imagine that our learning rule outputs a model q that always generates the same content z which is completely original and not similar to any protected content c. However, depending on the safe models q_c it can be the case that the model q is not similar to any of them.

The above differences, potentially, allow NAF to circumvent some of the hurdles for using DP as a notion for copyright. For example, the one-sidedness seems sufficient for copyright and may allow models that are discarded via DP. Also, the distinction between model-safety and content-safety can, for example, allow models that may memorize completely the training set as long as a content they output does not provide a proof for such memorization. Next, the fact that NAF is defined by a set of constraints, and not a property of the learning algorithm, allows one to treat breaches of Eq. (2) as soft "flagging" and not necessarily as hard constraints. This advantage is further discussed in Section 4. Finally, perhaps most distinguishable, is the possibility to use general safety functions that can capture copyright breaches more flexibly. We next discuss the implications of these refinements. We begin with the question of model safety vs. content safety in NAF and in DP.

Model safety vs. Content safety Our first result is a parallel to Theorem 3.1 in [38] in the context of DP stability. Theorem 3.1 in [38] shows how to efficiently transform a given learning rule A to a learning rule B which is NAF-stable, provided that A tends to output similar generative models when given inputs that are identically distributed. We state and prove a similar result by replacing NAF stability with DP stability.

Recall that the total variation distance between any two distributions is defined as:

$$\|q_1 - q_2\| = \frac{1}{2}\sum |q_1(x) - q_2(x)| = \sup_E \left(q_1(E) - q_2(E)\right),$$

Proposition 1. *Let A be an algorithm mapping samples S to models q_S^A such that*

$$\mathop{\mathbb{E}}_{S_1,S_2}\left[\|q_{S_1}^A - q_{S_2}^A\|\right] \le \alpha,$$

where $S_1, S_2 \sim D^m$ are two independent samples. Then, there exist an (ϵ, δ) DP algorithm B that receives a sample $S_B \sim D^{m_{priv}}$ such that if

$$m_{priv} = \tilde{O}\left(\frac{m}{\eta\epsilon}\log 1/\delta\right)$$

and $S_A \sim D^m$ then:

$$\mathop{\mathbb{E}}_{S_A, S_B}\left[\| \mathbb{E}[q^B_{S_B}] - q^A_{S_A}\|\right] \leq \frac{2\alpha}{1+\alpha} + O(\eta).$$

Where the expectation within, is taken over the randomness of B.

The premise in the above theorem is identical to that in Theorem 3.1 in [38] and captures the property that A provides similar outputs on identically distributed inputs. The obtained algorithm B is DP-stable and at the same time it has a similar functionality like A in sense that its output model q^B generates content z which in expectation is distributed like contents generated by q^A.

Safety functions We now turn to a discussion on the potential behind the use of different safety functions. The crucial point (which we discuss in great detail in Section 3 below) is that a satisfactory "copyright definition" *must* allow algorithms to be highly influenced, even by their input content which is *protected*. This reveals a stark contrast with algorithmic stability: it is easy to see that DP does not allow such influence. Indeed, the whole philosophy behind privacy is that a model is "safe" if it did not observe the private example (and in particular was not influenced by it).

This raises the question of whether the greater flexibility of the NAF model can provide better aligned notions of safety. In fact, if it is allowed to be influenced by protected data, one might even want to consider safe models that have *intentionally* observed a certain content and derived out of it the derivatives that are not protected.

The next result, though, shows that there is a *no free lunch* phenomenon. For every protected content c, we can either only consider safe models that observed c and are influenced by it, or only safe models that *never* observed it and were *not* influenced by it. In other words, if a protected content c influenced its safe model q_c then it must influence all safe models $q_{c'}$ for all protected contents c'. We further elaborate on the implication of this result in Section 4.

Below, q_1 and q_2 should be thought of as safe models, and p as the model outputted by the NAF learning algorithm. (So, in particular p should satisfy Eq. (2) w.r.t q_1 and q_2.) This result complements Theorem 3.1 in [38] which shows that NAF can be satisfied in the sharded-safety setting when the two safe models are close in total-variation.

Proposition 2. *Let q_1 and q_2 be two distributions such that*

$$\|q_1 - q_2\| \geq \alpha,$$

then for any distribution p we have that for some z:

$$p(z) \geq \frac{1}{2(1-\alpha)} \min\{q_1(z), q_2(z)\},$$

The proof is left to Appendix A.1.

3 The gap between algorithmic stability and copyright

So far, we provided a technical comparison between existing notions in the CS literature aimed towards provable copyright protection. While the technical notion for privacy may seem closely related, as observed through NAF there are differences and there is room for more refined definitions that may capture essential differences. While algorithmic stability approaches hold promise in helping courts assess copyright infringement cases (an issue we further discuss in Section 4), as we next discuss they cannot serve as a definitive test for copyright infringement. In order to see that, we next discuss the issue of copyright in the lens of the law. From a legal perspective, formal algorithmic stability approaches are both over and under exhaustive. Consequently, we will categorize this section based on these challenges.

3.1 Over-exhaustiveness

Here we focus on a concern that algorithmic stability approaches may filter out lawful output content that does not infringe copyright in the input content. Because non-infringing output content is lawful, employing algorithmic stability approaches as filters to generative models may needlessly limit their production capabilities, and, thereby, undermine the goals of copyright law.

Copyright law intends to foster the creation of original works of authorship by securing incentives to authors and, at the same time, ensuring the freedom of current and future authors to use and build upon existing works. The law derives from the U.S. Constitutional authority: "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." [5]

However, the goal of promoting progress is often inconsistent with unlimited rights to control copyrighted materials. For this reason, copyright law set fundamental limits on the rights it is granting to authors. Promoting progress is inconsistent with , because creators and creative processes are situated in cultural contexts. The creative process often involves an ongoing interaction with multiple stakeholders, rather than authorship in silos. Creating new works involves human capital enhanced by learning and research, engagement with pre-existing materials and a shared cultural language. Consequently, the use of copyrighted materials is an important input in any creative process. Such process often involves learning from pre-existing materials, applying existing styles, and referencing shared symbols and common works to generate a new interpretations and fresh meanings [18, 22].

For this reason, unlike the mandate of the algorithmic stability approaches, copyright law does not require that an output content will not draw at all on an input content to be lawful and non-infringing. On the contrary. There are many cases where copyright law explicitly allows for an output content to heavily draw on the input content without raising infringement concerns. In such cases, allowing an input content to impact an output content is not only something that copyright law permits, it is something that copyright law encourages. Doing so, as Jessica Litman put it: "is not parasitism; it is the essence of authorship." [31]

Copyright law allows an output content to substantially draw on an input content in three main cases, which we next list and then explore these categories:

1. When the input content is in the public domain.

2. When the input content is copyrighted but incorporates aspects that are excluded from copyright protection.

3. When the use of the protected aspects of the input content are lawful.

When an input content is in the public domain An input content may be unprotected because its copyright term has lapsed. Copyrights are limited in duration (though relatively a long duration, which in most countries will last life of the author plus seventy years). Once the copyright term expires, an input content enters the public domain and could freely be used and impact an output content without risking copyright infringement [31]. Public domain materials may also consist of anything that is not at all copyrightable, such as natural resources. For instance, if two photographers are taking a picture of the same person, some similarity between the pictures might be due the how this person looks, which is in the public domain. Other elements such as an original composition, or the choices made regarding lighting conditions and the exposure settings used in capturing the photograph may be considered copyrighted expression. If the generative model only makes use of the former in the output content, it may not constitute an infringement.

When an input content incorporates unprotected aspects Input contents with a valid copyright term, enjoys "full" legal protection, but it too is limited in scope. As provided by the copyright statute, "[i]n no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." [1] According to this principle, an output content may substantially draw on an input content without infringing the latter's copyrights, for as long as such taking is limited to the input's content unprotected elements.

- **Procedures, processes, systems and methods of operation** Copyright protection does not extend to "useful" or "functional" aspects of copyrighted works such as procedures, systems, and methods of operation. These aspects of an input content are freely accessible for an output content to draw upon. . For example, in the seminal case of Baker vs. Selden, the Supreme Court allowed Baker to create a book covering an improved book-keeping system while drawing heavily on the charts, examples, and descriptions used in Selden's book without infringing Selden's copyright [3]. As the court explained, the aspects that Baker took from Selden's work are functional methods of operations and as such are not within the domain of copyright law. Similarly, in Lotus v. Borland, the United States Court of Appeals for the First Circuit allowed Borland to copy Lotus's menu command hierarchy for its spreadsheet program, Lotus 1-2-3. The court ruled that Lotus menu command hierarchy was not copyrightable because they form methods of operation [8] - Consequently, if a generative model simply extracts procedures, processes, systems and methods from the training set it may not infringe copyright.

- **Ideas** Copyright protection is limited to concrete "expressions" and does not cover abstract "ideas." Thus, in Nicholas v. Universal, the United States Court of Appeals for the Second Circuit allowed Universal to incorporate many aspects of Anne Nichols' play Abie's Irish Rose, in their film The Cohens and Kellys [11]. The court explained that the narratives and characters that Universal used ("a quarrel between a Jewish and an Irish father, the marriage of their children, the birth of grandchildren and a reconciliation"), were "too generalized an abstraction from what she wrote. . . [and, as such]. . . only a part of her [unprotected] 'ideas.'" [11] When a generative model simply extract ideas from copyrighted materials, rather than replicating expressive content from their training data, it does not trigger copyright infringement.

- **Facts** Copyright protection also does not extend to facts. For example, in Nash v. C.B.S., the court ruled that C.B.S. could draw heavily from Jay Robert Nash's books without infringing his copyright [10]. As the court explained, the hypotheses that Nash rose speculating the capture of the gangster John Dillinger and the evidence he gathered (such as the physical differences between Dillinger and the corpse, the planted fingerprints, and photographs of Dillinger and other gangsters in the 1930s) were all unprotected facts that Nash could not legally appropriate. Consequently, generative models which simply memorize facts do not infringe copyright law.

When the use of the protected aspects of the input content was lawful Even when the protected elements of an input content ("expressions" rather than the "ideas") are impacting an output content, such impact may be legally permissible. There are two main categories of lawful uses: de minimis copying and fair use.

- **De minimis copying** Copyright law allows de minimis copying of protected expression, namely the coping of an insignificant amount that has no substantial impact on the rights of the copyright owner or their economic value. In a similar way, "[w]ords and short phrases, such as names, titles, and slogans, are uncopyrightable."[37]. However, de minimis coping of protected expression may be unlawful if it captures the heart of the work [4]. For example, phrases like "E.T. Phone Home." [12]
- **Fair Use** Copyright law also allows copying of protected expression if it qualifies as fair use. The U.S. fair use doctrine, as codified in § 107 of the U.S. Copyright Act of 1976, is yet another legal standard to carve out an exception for an otherwise infringing use after weighing a set of four statutory factors. The four statutory factors are: (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work [1].

Importantly, the fair use claimant need not satisfy each factor in order for the use to qualify as fair use [7]. Nor are the four factors meant to set out some kind of mathematical equation whereby, if at least three factors favor or disfavor fair use, that determines the result [33]. Rather, the factors serve as guidelines for holistic, case-by-case decision. In that vein, in its preamble paragraph, § 107 provides a list of several examples of the types of uses that can qualify as fair use. The examples, which include "criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, [and] research,"[1] are often thought to be favored uses for qualifying for fair use. Importantly, however, the list of favored uses is not dispositive. Rather, fair use's open-ended framework imposes no limits on the types of uses that courts may determine are "fair" [7].

When the factors strongly favour a finding of fair use, even output contents that are heavily impacted by copyrighted input contents may be excused from copyright infringement. For example, in Campbell v. Acuff-Rose, although the rap music group 2 Live Crew copied significant portions of lyrics and sound from Roy Orbison's familiar rock ballad "Oh, Pretty Woman" [7]. The Supreme Court denied liability in this case, based on the premise that the 2 Live Crew's derivative work was considered a "parody" of Orbison's original work, and, therefore, constituted fair use. Similarly, in The Authors Guild v. Google, the court defended Googles' mass digitization of millions of copyrighted books to create a searchable online database as fair use, because it considered Google's venture to be socially desirable [2] as explained by Sag [34] the copying of expressive works for non-expressive purposes should not be counted as a copyright infringement and must be considered fair use.

3.2 Under Exhaustiveness

Algorithmic stability approaches are under-exhaustive because they might fail to filter out unlawful output content that infringes copyright in the input content. As explained, algorithmic stability approaches find infringement only when the output content heavily draws on an input content. The law of copyright infringement, however, is not so narrow. Copyright law only requires that the output content will heavily draw on the protected expression that originated from an input content to find infringement. Such expression need not necessarily come from the input content itself; it may come from various other sources including copies, derivatives or snippets of the original content.

To illustrate this point, consider the fact pattern in the pending Supreme Court case Warhol vs. Goldsmith [6]. In that case, the portrait photographer Lynn Goldsmith accused Andy Warhol of infringing copyrights in a photograph she took of the American singer Prince. Goldsmith consented to Warhol to use her photograph as an "artistic reference" for creating one derivative illustration[1]. Still, she did not approve nor imagine that Warhol had, in fact, made 16 different derivatives from the original photograph. Warhol's collection of Prince portraits, also known as the Prince series , is depicted in Fig. 1, right side.

For our purposes, assume the Prince Series' portraits served as input for a generative machine. If the machine's output content draws heavily on Goldsmith's protected expression that is baked into the Prince Series' portraits, then the output content may infringe Goldsmith's copyright in original photograph (Fig. 1 , left side), even if the machine did not have access to Goldsmith's original photograph. Moreover, this risk will not be eliminated even if the Supreme Court ends up deciding that the Prince Series' portraits themselves are non-infringing because they constitute fair use.

Putting it simply, copying from a derivative work—whether authorized by the copyright owner or not— may infringe copyright in the original work on which the derivative work is based. This situation is prevalent in copyright practice, especially in music. In modern music copyright cases, plaintiffs usually show access to the original copyrighted work (musical composition) by showing access to a derivative work of that original work (sound recording). Plaintiffs are not required to demonstrate that the defendants also had access to the original sheet music nor that they could actually read musical notes.

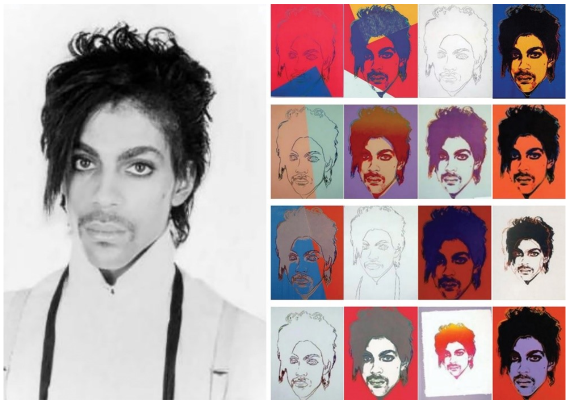

Figure 1

Lastly, output content can also infringe copyright in input content by accessing parts or snippets of the input content even without accessing the input content in its entirety. This concern was raised recently in The Authors Guild v. Google, a case dealing with the legality of the Google Book Search Library Partner project [2]. As part of this project, Google scanned and entered many copyrighted books into their searchable database but only provided "snippet views" of the scanned pages in search results to their users. The plaintiff in the case argued that Google facilitated copyright infringement by allowing users to aggregate different snippets and reconstruct infringing copies of their original works. The court ended up dismissing this claim, but only because Google took affirmative steps to prevent such reconstruction by limiting the number of available snippets and by blacklisting certain pages.

To sum up, there are numerous instances where copyright law allows (even encourage) an output content to draw on an input content. The more substantial are the unprotected aspects an input content, and the more likely it is that using the input content's protectable aspects is considered lawful, the more expansively can the output content draw on the input content without fearing

[1] bottom right most picture in Fig. 1

copyright infringement.

4 Discussion

Algorithmic stability approaches establishing a proof of copyright infringement are either too strict or too lenient from a legal perspective. Due to this misfit, applying algorithmic stability approaches as filters for generative models is likely to distort the delicate balance that copyright law aims to achieve between economic incentives and access to creative works.

This is not to say that algorithmic approaches in general and algorithmic stability approaches in particular has no value to the legal profession. Quite the opposite. Computer science methodologies bring a significant benefit to the judicial table: the ability to weigh large volumes of information and assist policymakers in making more informed decisions. Rather than constructing binary legal rules (e.g., a definitive test for copyright infringement), algorithmic approaches should facilitate new quantifiable methods for applying legal standards. Such methods could help clarify vague legal concepts such as "fairness," "privacy," and, in the copyright context—"originality", and at the same time facilitate the ongoing development of legal and social norms.

Stability is not safe Nevertheless, it is also important to understand the exact limitations of such algorithmic approaches. The NAF framework, that allow a rich class of safety functions, has the potential to circumvent some of the challenges presented, but may still be limited and we now wish to discuss this in further details.

To utilize the NAF framework, the first basic question one needs to address, when using the NAF framework, is

Given a protected content c how should we choose the safe model $\text{safe}(c)$?

It seems natural to include models that are not heavily influenced by c since otherwise this might allow copyright breaching. However, such choice of $\text{safe}(c)$ leads to the limitations encountered by the DP setting. It is true that some aspects, such as content safety vs. model safety, can be better aligned through the definition of NAF but also, as Proposition 1 shows, through variants of DP. Overall, there is room, then, to further investigate the different possible models for copyright, within such an approach, but we should take into account the limitations presented in Section 3.

Perhaps a more exciting application of NAF, then, is to consider notions of safety that allow some influence by c. e.g. to enable generating parodies, fair-use, de minimis copying, etc. We consider then safety functions that now *do* have access to c, and exploit this access to validate that only allowed influence happens. Here we face a different challenge. Suppose that q_c is such a safe model for content c. Suppose, also, that $q_{c'}$ is another safe model for content c'. If $q_{c'}$ and q_c are far away, then Proposition 1 shows that there is no hope to output a NAF model. But even if q_c and $q_{c'}$ are not far away, but suppose that $q_{c'}$ ignores content c, then for any content z that is influenced by c we may assume that:

$$q_c(z) \gg q_{c'}(z).$$

But, if p is a NAF model, we must also have due to Eq. (2) with respect to c' and z:

$$q_c(z) \gg p(z).$$

In other words, the NAF model censors permissible content z even though it is safe. This happens because z is an improbable event in model $q_{c'}$. Not because z breaches copyright of c' but because it is influenced by c, and content that is influenced by c is discarded by safe models that had no access to c.

It follows, then, that all safe models must treat protected content in a similar manner, and $q_{c'}$ must also be influenced by c if we expect the NAF model to make any use of it. Hence, it is unclear if a more refined notion of safe may help circumvent the hurdles of the privacy approach. This suggests, though, to perhaps consider a relaxed variant of NAF in which a content is discarded by a safe model only when certain links between the protected content and the generated content are established.

It seems, then, that an algorithmic approach that assist jurists in understanding such links between existing works of authorship, study their hidden interconnection, and quantify their originality can hold great promise. From this perspective, originality is evaluated by the semantic distance between a measured expressive work and similar materials found in the corpus of pre-existing expressions. Research in this area is just beginning but holds a great promise for the copyright system.

Ackgnoweledgments We thank Bruria Friedman for research assistance.

References

[1] 17 U.S.C. § 102(b) (2006). 8, 9

[2] Authors Guild v. Google, Inc., 804 F.3d 202, 207–08, 225 (2d Cir. 2015)). 2, 9, 10

[3] Baker v. Selden, 101 U.S. 99 (1879). 8

[4] Harper & Row v. Nation Enterprises, 471 U.S. 539 (1985). 9

[5] U.S. CONST. art. I, § 8, cl. 8 . 2, 7

[6] Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith (Docket 21–869)]. . 10

[7] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 578 (1994). . 9

[8] Lotus Dev. Corp. v. Borland Int'l, Inc., 49 F.3d 807, 815 (1st Cir. 1995); Lotus Dev. Corp. v. Borland Int'l, Inc., 516 U.S. 233 (1996) . 8

[9] Mazer v. Stein, 347 U.S. 201, 219 (1954)] . 2

[10] Nash v. CBS, Inc., 899 F.2d 1537 (7th, cir., 1990). 8

[11] Nichols v. Universal Pictures Corporation, 45 F.2d 119, (2st Cir., 1930) . 8

[12] Universal City Studios v. Kamar Industries, Inc., 217 U.S.P.Q. (BNA) 1165 (S.D. Tex 1982). 9

[13] O. Angel and Y. Spinka. Pairwise optimal coupling of multiple random variables. *arXiv preprint arXiv:1903.00632*, 2019. 14

[14] C. D. Asay. Independent creation in a world of ai. *FIU L. Rev.*, 14:201, 2020. 3

[15] O. Bousquet, R. Livni, and S. Moran. Synthetic data generators–sequential and private. *Advances in Neural Information Processing Systems*, 33:7114–7124, 2020. 2, 3

[16] M. Bun, K. Nissim, and U. Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 369–380, 2016. 14

[17] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 2, 3

[18] J. E. Cohen. *Configuring the networked self: Law, code, and the play of everyday practice.* Yale University Press, 2012. 7

[19] C. Dwork and V. Feldman. Privacy-preserving prediction. In *Conference On Learning Theory*, pages 1693–1702. PMLR, 2018. 3, 5

[20] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006. 4

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 3

[22] N. Elkin-Koren. Cyberlaw and social change: A democratic approach to copyright law in cyberspace. *Cardozo Arts & Ent. LJ*, 14:215, 1996. 7

[23] G. Franceschelli and M. Musolesi. Deepcreativity: measuring creativity with deep learning techniques. *Intelligenza Artificiale*, 16(2):151–163, 2022. 3

[24] J. Grimmelmann. Copyright for literate robots. *Iowa L. Rev.*, 101:657, 2015. 2, 3

[25] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani. Reconstructing training data from trained neural networks. *arXiv preprint arXiv:2206.07758*, 2022. 3

[26] P. Henderson, X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023. 3

[27] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. 3

[28] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pages 171–180, 2009. 14

[29] Legislation and L. C. C. Law). *OPINION: USES OF COPYRIGHTED MATERIALS FOR MACHINE LEARNING.* State of Israel Ministry of Justice, 2022. 2

[30] M. A. Lemley and B. Casey. Fair learning. *Tex. L. Rev.*, 99:743, 2020. 2, 3

[31] J. Litman. The public domain. *Emory Lj*, 39:965, 1990. 2, 7

[32] N. W. Netanel. *Copyright's paradox.* Oxford University Press, 2008. 3

[33] N. W. Netanel. Making sense of fair use. *Lewis & Clark L. Rev.*, 15:715, 2011. 2, 9

[34] M. Sag. The new legal landscape for text mining and machine learning. *J. Copyright Soc'y USA*, 66:291, 2018. 9

[35] P. Samuelson. Reconceptualizing copyright's merger doctrine. *J. Copyright Soc'y USA*, 63: 417, 2016. 2

[36] S. Scheffler, E. Tromer, and M. Varia. Formalizing human ingenuity: A quantitative framework for coyright law's substantial similarity. *arXiv preprint arXiv:2206.01230*, 2022. 2, 3

[37] C. . U.S. Copyright Office. Works Not Protected by Copyright. 2021. 9

[38] N. Vyas, S. Kakade, and B. Barak. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023. 2, 3, 4, 5, 6

A Proofs

A.1 Proof of Proposition 2

Suppose that

$$\|q_1 - q_2\| \geq \alpha.$$

In particular there exists an event E such that:

$$q_2(E) \leq q_1(E) - \alpha \leq 1 - \alpha.$$

Let p be some distribution. We assume that $p(E) \geq 1/2$ (otherwise, replace E with its complement and q_1 and q_2 replace roles). Thus, we have that:

$$p(E) \geq \frac{1}{2} \geq \frac{1}{2(1-\alpha)} q_2(E).$$

In particular, for some $z \in E$, the result follows.

A.2 Proof of Proposition 1

The proof relies on a coupling Lemma, taken from [13]. Recall that, given a collection of distribution measures Q, a coupling can be thought of as a collection of random variables $X = (X_q)_{q \in Q}$, whose marginal distributions are given by q. I.e. $\mathbb{P}(X_q = x) = q(x)$:

Lemma 1 (A special case of Thm 2 in [13]). *Let Q be the collection of all posteriors over a finite domain $\mathcal{X}$[2]. There exists a coupling such that for every $q, q' \in Q$:*

$$\mathbb{P}(X_q \neq X_{q'}) \leq \frac{2\|q - q'\|}{1 + \|q - q'\|}.$$

The second Lemma we rely on is a private heavy hitter mechanism, described as follows:

Lemma 2 ([16, 28]). *Let Z be a finite data domain. For some*

$$k \geq \Omega\left(\frac{\log 1/\eta\beta\delta}{\eta\epsilon}\right),$$

[2] which are all absolutely continuous w.r.t the uniform distribution

there exists an (ϵ, δ)-DP algorithm hist*, such that with probability $(1-\beta)$ on an inputs $S = \{z_1, \ldots, z_k\}$ outputs a mapping $a \in [0,1]^Z$, such that, for every $z \in Z$,*

$$|a(z) - freq_S(z)| \leq \eta.$$

In particular, if $freq_S(z) > 0$, then $a(z) > 0$.

Where we denote by $\text{freq}_S(z) = \frac{|i:z_i=z|}{|S|}$.

We next move on to prove the claim. Let X be the coupling from Lemma 1. Our private algorithm works as follows:

1. First, we take $\beta = \eta$, and set
$$k = \Omega\left(\frac{\log 1/\eta^2\delta}{\eta\epsilon}\right).$$
To be as in Lemma 2.

2. Divide S, the input sample, to k, disjoint datasets $S_1, \ldots, S_k$ of size m. Each data set, via A, defines a model $q^A_{S_i}$.

3. Next, we define the random sample
$$S_X = \{X_{q^A_{S_1}}, X_{q^A_{S_2}}, \ldots, X_{q^A_{S_k}}\} \in Z^K.$$

4. Apply the mechanism in Lemma 2 and output $a \in [0,1]^Z$ such that, w.p. $1-\eta$, for all $z \in Z$:
$$|a(z) - \text{freq}_{S_X}(z)| \leq \eta.$$

5. Let p be any arbitrary distribution such that for every $z \in Z$:
$$|a(z) - p(z)| \leq \eta \tag{3}$$
(if no such distribution exists p is any distribution). and output
$$q^B_S = p.$$

Notice that each sample z_j affects only a single sub-sample S_i and in turn only a single random variable $X_{q^A_{S_i}}$. The histogram function a is then (ϵ, δ)-DP w.r.t to its input S. The output p, by processing is also private. We obtain, then, that the above algorithm is (ϵ, δ)-private.

We next set out to prove that $p = q^B_S$ is close in TV distance to $q^A_{S_A}$ in expectation. For ease of notation let us denote $X_i = X_{q^A_{S_i}}$. Notice that, with probability $(1-\eta)$, for every z:

$$|a(z) - \text{freq}_{S_X}(z)| \leq \eta,$$

in particular, there is a p that satisfies the requirement in Item 5 (i.e. freq_{S_X} defines such a distribution) and Eq. (3) is satisfied. We then have that for every z:

$$\begin{aligned}\left|p(z) - \frac{1}{k}\sum \mathbf{1}[X_i = z]\right| &\leq |p(z) - a(z)| + \left|a(z) - \frac{1}{k}\sum \mathbf{1}[X_i = z]\right| \\ &\leq 2\eta. \end{aligned} \tag{4}$$

We now move on to bound the total variation between the model $\mathbb{E}[q_S^B]$ and q_{S_A}, where expectation is taken over the randomness of B.

To show this, we will use the reverse inequality of the coupling Lemma, in particular if $(\hat{X}_B, \hat{X}_A)$ is a coupling of q_S^B and $q_{S_A}^A$ (where S and S_A are now fixed), then:

$$\| \mathbb{E}[q_S^B] - q_{S_A}^A \| \leq \mathbb{P}(\hat{X}_B \neq \hat{X}_A). \tag{5}$$

Our coupling will work as follows, first we output $p = q_S^B$ and sample $\hat{X}_B \sim p$, and we let $\hat{X}_A = X_{q_{S_A}}$. This defines a coupling $(\hat{X}_B, \hat{X}_A)$. Applying Eq. (4), with $z = \hat{X}_A$, exploiting the fact that Eq. (4) holds with probability at least $1 - \eta$:

$$\begin{aligned} \mathbb{P}(\hat{X}_B \neq \hat{X}_A) &\leq \frac{1}{k} \sum_{i=1}^{k} \mathbb{P}(X_i \neq X_{q_{S_A}}) + \eta \\ &\leq 2\eta + \eta. \end{aligned}$$

And we have that:

$$\mathbb{P}(\hat{X}_B \neq \hat{X}_A) \leq \frac{1}{k} \sum_{i=1}^{k} \mathbb{P}(X_i \neq X_{q_{S_A}}) + 3\eta \leq \frac{1}{k} \sum_{i=1}^{k} \frac{2\|q_{S_i}^A - q_{S_A}\|}{1 + \|q_{S_i}^A - q_{S_A}\|} + 3\eta.$$

And,

$$\begin{aligned} \mathop{\mathbb{E}}_{S_A,S} \| \mathbb{E}[q_S^B] - q_{S_A} \| &\leq \mathop{\mathbb{E}}_{S_A,S} \frac{1}{k} \sum_{i=1}^{k} \left[\frac{2\|q_{S_i}^A - q_{S_A}\|}{1 + \|q_{S_i}^A - q_{S_A}\|} \right] + 3\eta \\ &\leq \mathop{\mathbb{E}}_{S_1,S_2 \sim S} \left[\frac{2\|q_{S_1}^A - q_{S_2}\|}{1 + \|q_{S_1}^A - q_{S_2}\|} \right] + 3\eta \\ &\leq \left[\frac{2\,\mathbb{E}[\|q_{S_1}^A - q_{S_2}\|]}{1 + \mathbb{E}[\|q_{S_1}^A - q_{S_2}]\|} \right] + 3\eta & \text{concavitiy of } \frac{2x}{1+x} \\ &\leq \left[\frac{2\alpha}{1+\alpha} \right] + 3\eta & \text{monotinicity } \frac{2x}{1+x} \end{aligned}$$