UCSD UNDERGRADUATE LAW REVIEW

ELAINE (LINGHUI) ZHAO

Overfitting and Copyright Infringement: How Should Copyright Law Address The Italian Plumber Problem?

ABSTRACT. This Article explores a discrete yet consequential driver of AI-generated copyright infringement: overfitting. By analyzing overfitting as a statistical phenomenon, it offers a legal framework for understanding how this technical failure can lead machine learning models to reproduce protected works nearly identically. Through the lens of the Italian Plumber Problem, this article critiques the limitations of the fair use doctrine when applied to generative AI outputs, identifies regulatory and policy shortcomings, and argues for an updated doctrinal toolkit. Finally, it draws a conceptual parallel between algorithmic overfitting and judicial reasoning, positing that courts may fall prey to overfitting by relying on fact-specific precedent in the absence of guiding rules for novel technologies. Using interdisciplinary analysis, this article calls for a recalibration of legal frameworks to accommodate the realities of machine learning and protect both innovation and authorship in the digital age.

AUTHOR. Elaine Zhao is a rising fourth-year at UC San Diego, double majoring in Political Science and Applied Mathematics. She is interested in the intersection of technology and law, particularly how existing copyright doctrines can evolve to address the challenges posed by generative AI. Her research explores the implications of model training on copyright liability, and she hopes to contribute to a more coherent framework for evaluating algorithmic outputs. She will be applying to law school in the upcoming cycle to further pursue public interest technology law and policy.

INTRODUCTION

The human mind was traditionally viewed as the only entity capable of learning and creativity, but recent advancements in Artificial Intelligence have challenged this notion. AI systems, such as ChatGPT and Stable Diffusion, now demonstrate capabilities previously considered exclusive to human creators. This development necessitates judicial consideration to address the legal implications arising from AI-generated outputs.

A legal implication accompanies each revolutionary advancement in technology. Several scholars have raised concerns about AI's potential copyright infringement, from the unauthorized use of copyrighted materials as training data and testing processes to AI-generated outputs that infringe copyright.[1] Because this issue involves advanced technology, it cannot be fully understood from only a legal perspective. We also need to explain how AI systems work from a technical point of view–including concepts like model training, machine learning algorithms, and regression analysis. AI modeling is a vast and complex field, and skipping technical details oversimplifies the problem. While some papers have explored AI's infringement of input training data, they are limited to the conceptual explanation of AI technology. At the intersection of technology and law, it is crucial to understand the technical mechanism behind overfitting before defending or making judgments.

This paper focuses on overfitting, one aspect of AI's potential copyright infringement arising from model training. I will dissect this phenomenon, technically exploring different root causes. The paper will introduce the Italian Plumber Problem, created by Professor Matthew Sag,[2] and examine how this example resulted in infringement due to overfitting. Subsequently, I will evaluate the fair use doctrine as a framework for AI copyright infringement cases and explore its unpredictability. This paper will conclude by proposing alternative approaches that offer more viable solutions. Finally, in the further discussion section, I will highlight an intriguing parallel between model development and judicial rulings, suggesting that overfitting is precisely why fair use analysis is not a good fit for AI's copyright infringement issues.

---

[1] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 53-54 (2017).; David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. Davis L. Rev. 653 (2017).

[2] Matthew Sag, *Copyright Safety for Generative A*I, 61 Hous. L. Rev. 295 (2023).

I. THE STATISTICAL PHENOMENON OF OVERFITTING

A. Concept of Overfitting: A Problematic Perfection

When training a model on a limited dataset, data scientists develop algorithms that enable the model to recognize underlying patterns and accurately represent relationships within the data. The presence of outliers and noise skew produces irrelevant factors to the core characteristics, complicating the data collection process. Outliers lie outside of "normal" data due to measurement error, data-entry mistakes, instrument malfunction, or genuine but rare variability in the underlying phenomenon.[3] Noise refers to random data patterns that are irrelevant to the traits of a subject.[4]

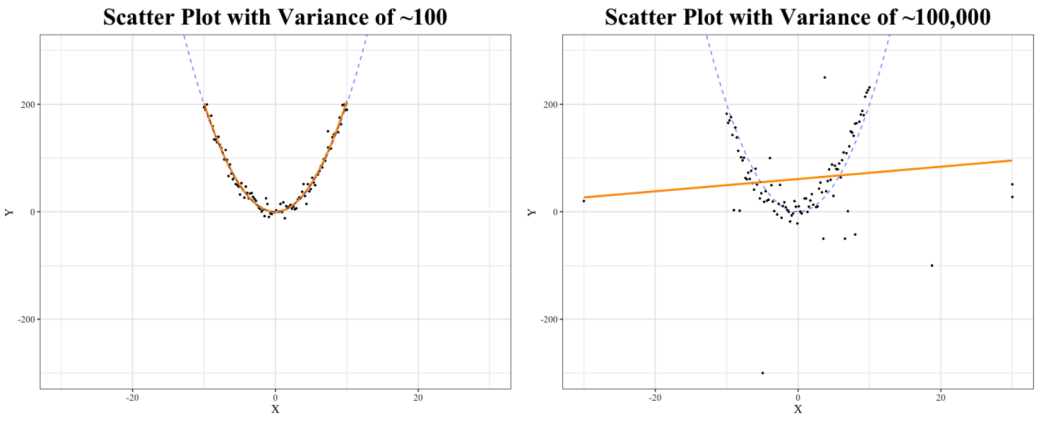

Figure 1 & 2: Datasets with Variance of ~100 and 100,000

To illustrate this, one considers a dataset in which the data points follow a simple quadratic equation $y = 2x^2$ with a low variance. For comparison, another dataset with higher variance is created, and an R code will be tasked with performing quadratic regression on both datasets. The figures above are clear examples. Figure 1 (on the left side) models a dataset with added random noise that creates a variance of ~100. This is the square of the vertical distance from each data point to the equation $y = 2x^2$.

[3] *What Are Outliers in the Data?*, Nat'l Inst. of Standards & Tech., https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm (last visited Aug. 8, 2024).
[4] Armin Bunde, *The Different Types of Noise and How They Effect Data Analysis*, 95 Chemie Ingenieur Technik 1758-1767 (2023).

Figure 2 (on the right side) has the same rules applied, except the data is twisted to create a significantly larger variance of ~100,000.[5] The equation $y = 2x^2$ is represented by a blue dotted parabola to provide a reference for a perfect fit. The orange solid line is the quadratic model regressed upon the given dataset. In the first dataset, where most data closely fits the equation and no extreme outliers are present, the algorithm can almost perfectly align with the desired equation.[6] However, in practical cases, noise and outliers obscure the underlying desired pattern. In Figure 2, the quadratic model has almost become a linear equation, which deviates significantly from the desired result.[7] The undifferentiated weight placed on noise and valid data points causes the model to overfit to the random variance, generating an undesired regression result far from the "truth." This leads to poor performance on predictive tasks.

A model that performs perfectly on training data and poorly at generating predictions on new data is a strong indicator of overfitting. When a model fits data such that it captures useful information while also keeping noise or outliers, it skews the output, making it inaccurate. To illustrate this, take a very simple analogy of a student preparing for a math quiz. Considering the following 4 questions:

1. An egg costs 1 dollar. How much does it cost to buy 2 dozen eggs?
2. An egg costs 3 dollars. How much does it cost to buy 3 dozen eggs?
3. A button costs 1 euro. How much does it cost to buy 3 dozen buttons?
4. I want to buy 2 dozen eggs. How many dollars do I need to complete the purchase if each egg costs 1 dollar?

The teacher is testing the students' understanding of the linear relationship between the price of a good and its quantity, as well as their understanding of multiplication. Assume the students were given the first question and answer. If a student understands the reasoning behind the first question, they would have no trouble completing the other three unseen questions that, although worded differently, test the same logic. If, instead of comprehending the concepts and theorems, a student only studies the specific wording and the format of the first question, they could confidently give the right answer when tested on the same question. However, if they are unable to comprehend that some parts of the problem (such as "egg," "dollar," and

[5] The more exact variances are 100.2033 and 99999.55.

[6] The first regressed model is approximately y = 2.04*x^2 + 0.21*x - 0.19.

[7] The second regressed model is approximately y = 0.0001*x^2 + 1.15*x +61.00.

the numbers used) are not relevant to the conceptual understanding of this linear relationship, a similar question would be difficult for the student to answer on future tests. Once the teacher changes the numbers or wording in the questions, replacing "egg" with "button" or "dollar" with "euro," the student will not be able to answer because their knowledge is deeply fit in the format rather than the underlying meaning of the question. To summarize, overfitting oversimplifies training data by creating a problematic "perfect fit" that cannot be applied accurately to similar problems.

Overfitted models sometimes result in memorization, the act of generating verbatim copies of training data. This is of particular concern to legal scholars.[8] When an AI model is the product of copyrighted work derived largely from memorization, infringement naturally follows. It is crucial to discern memorization from the concept of overfitting. Overfitting occurs when a model fits training data too closely and fails to generalize, performing poorly on unseen test data and making inaccurate predictions. Memorization focuses on the procedural act by which the computer remembers specific training data instead of making predictions.[9] This raw recall mechanism creates a direct pathway for expression to be copied word-for-word, triggering infringement irrespective of the model's generalization performance.[10]

B. When Does It Occur?

Developers have identified four common issues that lead to overfitting. Two causes arise from unrepresentative data, and two from inappropriate model design and training. This section will delve into each scenario, providing explanations for each cause.

[8] Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295 (2023).

[9] Layne Sadler, *Memorization isn't learning, it's overfitting*, Medium, Mar. 6, 2022, https://aiqc.medium.com/memorization-isnt-learning-it-s-overfitting-b3163fe6a8b4.

[10] It is worth noting that memorization is not directly related to copyright infringement: Infringement does not mean the computer makes a prediction and throws out what it was trained on. A model can output a perfect enough result that coincides with the expression of copyrighted work, even though the model never even has access to the original expressive data. There could be infringement risks not only on the output side but also the input of training data.

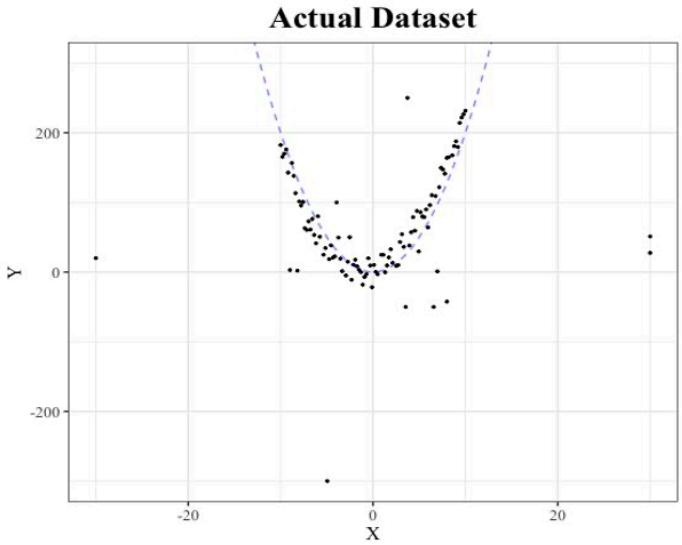

Figure 3: Scatter Plot of the Actual Dataset

I will use an example throughout this section to assist with discerning between the four issues leading to overfitting: noisy data, insufficient training data, overly complex models, and excessive training time. Assume we have a dataset of 100 data points that largely follows a quadratic relationship with certain variance, as shown in the figure. Five extreme outliers and certain random noises are introduced deliberately to demonstrate the effect of noise on prediction performance. A model is designed to regress on the data to fit most accurately (which would be the quadratic relation $y = 2x^2$). For each scenario, we plot the resulting regression line over the modified data to see how that particular factor skews the fit and impairs the model's ability to generalize new inputs.

1. Noisy Data

Noisy data is a large contributor to the problem of overfitting. When the proportion of random, non-informative variations in the training set grows large relative to the true underlying signal, the learning algorithm will "learn" these fluctuations as if they were genuine patterns. The patterns identified by the machine are often obscured by outliers or noise, especially when the dataset is limited or skewed. In this case, random errors, outliers, or irrelevant fluctuations make up such a large share of the data that the true signal is obscured, typically when the variance from noise exceeds the variance of the underlying pattern (e.g., noise accounts for 30–50 percent

or more of total variance). Learning algorithms will "learn" the noise instead of the real relationships, leading to overfitting. The noisiness could come from measurement error, design error, or an unrepresentative dataset. When a model is not designed to generalize datasets, random fluctuations are kept. High-capacity models with many parameters aiming to minimize training loss can adapt to every idiosyncrasy in the training set. Without proper constraints or bias toward simplicity, the optimization process treats noise as a genuine pattern, which degrades performance on unseen inputs.

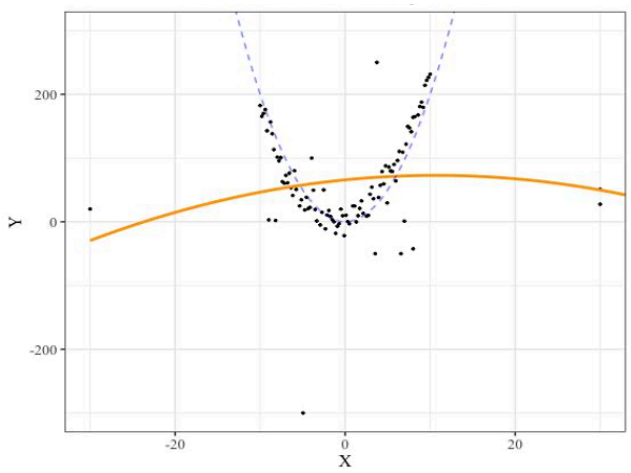

Figure 4: Overfitting due to Noisy Data

In the regression example, as shown in Figure 4, although the general trend follows a quadratic relationship, some data points deviate from the desired quadratic line, contributing to the variance. When a model is not designed to ignore such deviation, it will try to fit the data as closely as possible. The quadratic regression results in a negative, downward-facing parabola, while the actual quadratic coefficient is positive. For example, if a model is learning to identify the species of cat with pictures of cats that comprise the input, the machine would learn that a cat has two ears, two eyes, and fur, which are indeed accurate characteristics of a cat. However, random pictures, such as cats with bowties or sunglasses, might also be included in the dataset.

The model would not recognize a "cat" as defined by humans, unable to identify that the bowtie is an irrelevant factor. Instead, the model might recognize the bowtie as a trait that is inseparable from the species of cat, making false predictions when tasked with describing a cat or generating a picture. When trained on data unrepresentative of

the general reality, the model is unable to differentiate the outlier, noise, or meaningful patterns from the data, leading to inaccurate predictions.

2. Insufficient Training Data

The real world is infinitely varied, and practical limits (time, cost, privacy, the curse of dimensionality) make it difficult to perfectly capture every scenario. The model may not identify the pattern behind the dataset if the data is too limited to demonstrate a reliable trend. Using the regression example, if the model is provided with deficient data points, two problems can arise. In this example, we already know the right fit should be $y = 2x^2$, and we can instruct the algorithm to regress on quadratic relationships. In reality, neither the AI developers nor the AI model always know the correct underlying patterns behind the data. Insufficient data could cause the model to apply the wrong type of regression, resulting in poor pattern prediction.

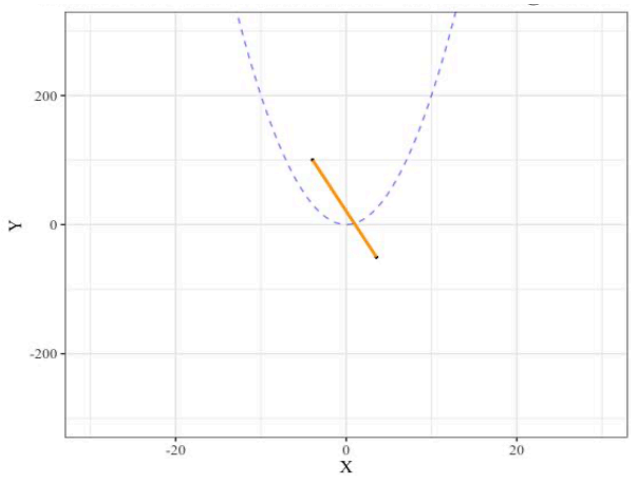

Figure 5: Overfitting due to insufficient Data - Wrong Regression Choice

Figure 5 illustrates this idea with an extreme example. When only two data points are made available, the most straightforward connection between the two points is the linear line passing through them, even though the actual regression should be quadratic.

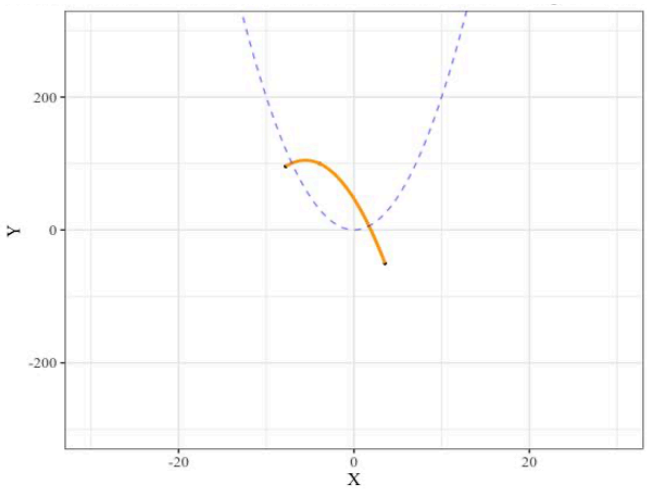

Figure 6: Overfitting due to insufficient Data - Skewed data

Even if the model made a successful choice on the type of regression, noise and outliers may still affect the model. The data point in practice will not perfectly follow along $y = 2x^2$. Figure 8 demonstrates that when regressing on three data points only, the result substantially diverges from the correct answer due to the presence of unrepresentative, noisy data points. When there is less data available, discerning between noise and the underlying pattern becomes much more difficult. More weight is then placed on random variance in the limited dataset, exacerbating the influence of noise and skewing the final result.

This proves that data shortages can yield problematic results. If available information is too finite, when abstracting a general pattern for making predictions, the model might choose to memorize the dataset. The input is so specific that the model does not have other available information to "group" them with.

3. Overly Complex Model

Complex models that include too many model parameters, such as CNN or

RNN,[11] may cause the model to catch noise and specific patterns in the training data.[12] Figure 9 illustrates a scenario where a 10-degree polynomial model is fit into a dataset. The dataset is designed to follow a quadratic trend with the curve concave up. Fitting a higher-degree function to the lower-degree function would not produce the desired output, even if the model assumes the 10-degree polynomial has higher precision. This is due to the model being too complex, as it would have many parameters subject to fluctuations within different minor waves in the model. This would sacrifice the accuracy of the output, even though it fits the available training data. In the figure below, the orange solid curve twists inward and concaves multiple times to fit the extreme outliers on the two ends of the X-axis. Generality pattern abstraction is conceded to overfit noisy data points. I propose that the data analysts and AI developers must trade off between overfitting and underfitting to achieve optimal generalization performance by carefully balancing model complexity and data representation.

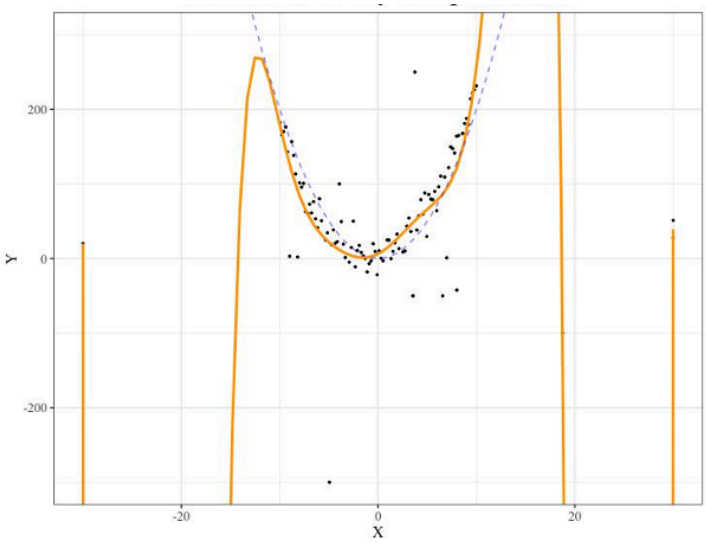

Figure 9: Overfitting due to an Overly Complex Model

[11] Javier Herbas, *Deep Learning Neural Network: Complex vs. Simple Model*, Medium, Oct. 10, 2020, https://medium.com/analytics-vidhya/deep-learning-neural-network-complex-vs-simple-model-88f6dcf88eaa.

[12] Mayank Mishra, *Convolutional Neural Networks, Explained*, Towards Data Science, Aug. 26, 2020, https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939/?gi=f1eacf1fb27e.

4. Excessive Training Time

To understand the final issue, it is crucial to understand the idea of epoch, or the number of times a dataset is passed to a model for learning in a cycle.[13] The epoch allows the model to refine its weight assignments and improve accuracy. With this and an increase in training time, the model gains better knowledge of the set and fits closer to the training data. This helps avoid underfitting a model to the training set, reflected in a reduction of errors on training data tests. However, excessive training time may cause the model to gradually adapt to specific patterns in the training data, neglecting the model's generalization capability. In this case, instead of underfitting, overfitting occurs because when a model is repeatedly studied, it results in the capture of noise or outliers.

Under this context, it is interesting to examine the high-profile lawsuit that is still pending: *The New York Times Company v. Microsoft Corporation*.[14] The New York Times is suing Microsoft and its minority partner OpenAI, the creator of ChatGPT, for infringing upon the expression of their articles.[15] Microsoft is accused of taking work produced by the New York Times and using it as training data to generate near-perfect output of the news articles' content. Showcased in Exhibit J of the Plaintiff's complaint, when Chat-GPT 4.0 was tasked with completing the second half of a copyrighted article published by the New York Times, it generated a near-exact copy of the second half of the original article.[16] This demonstrates that the Chat-GPT 4.0 model was likely memorizing content from the New York Times article.[17] OpenAI responded by accusing the plaintiff of hiring a hacker to deliberately induce infringing output and requested a case dismissal. The dismissal is based upon the claim that the New York Times made "tens of thousands of attempts" to generate the anomalous

[13] Aditya Kumar, *What is Epoch in Machine Learning?*, Simplilearn, Apr. 2, 2025, https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-epoch-in-machine-learning.
[14] The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023).
[15] *Id.*
[16] Ex. J to Compl., The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023), https://storage.courtlistener.com/recap/gov.uscourts.nysd.612697/gov.uscourts.nysd.612697.1.68.pdf.
[17] *One Hundred Examples of GPT-4 Memorizing Content from the New York Times*, N.Y. Times, Dec. 27, 2023, https://storage.courtlistener.com/recap/gov.uscourts.nysd.612697/gov.uscourts.nysd.612697.1.68.pdf.

result in the exhibits,[18] which is not what normal users would do when utilizing ChatGPT. If OpenAI's contention is true, the New York Times may have exploited a weakness common in machine learning and model training. Repeatedly feeding the model identical copyrighted texts, the New York Times increased the epochs of its articles. Since ChatGPT is exposed to the same specific expressions, it overfitted to the data and generated nearly indistinguishable output. This undesired overfitting resulted in data regurgitation, as the model memorized the input data. If the courts rule in favor of the New York Times, it could force AI developers to license or filter copyrighted inputs, raising engineering and compliance costs.

C. Techniques for Detection: How to Identify Overfitting?

Under this subsection, I will present common techniques used by programmers to detect the presence of overfitting. Since overfitting refers to a model failing to make accurate predictions on unencountered data, the testing mechanism takes part of the available training data as a test group to see how the model behaves. Poor performance on new data often signals that a model has overfitted to its training set–memorizing noise rather than learning patterns. To ensure a model remains robust and generalizable, three common detection techniques can be employed:

1. Holdout Validation

Holdout Validation is the simplest validation technique that data scientists use to detect overfitting.[19] It involves splitting the available data into two parts: the training set and the testing set. A great discrepancy between the performance of the two sets suggests that the model is overfitted to the data set, whereas a small discrepancy indicates that it generalizes well to unseen data, capturing genuine patterns rather than memorizing noise.[20]

2. K-fold Cross-Validation

---

[18] Robert Freedman, *OpenAI seeks to trim NYT lawsuit to fair use question*, Legal Dive, Feb. 28, 2024, https://www.legaldive.com/news/openai-seeks-trimming-nyt-infringement-lawsuit-to-fair-use-question-copyright-law-chatgpt/708805/.
[19] *Introduction of Holdout Method*, GeeksforGeeks, Aug. 26, 2020, https://www.geeksforgeeks.org/introduction-of-holdout-method/.
[20] *Id.*

In a cross-validation testing method, data scientists break down the training data into equally sized subsets (K) called folds.[21] A series of iterations will start with (K-1) folds being fed into the learning model, and the leftover fold (K) used as the "validation data" to test out the model's performance.[22] The iteration will not stop until all subsets have been tested and an average score is returned. This provides a reliable estimate of the model's generalization performance and guides any necessary hyperparameter tuning, which refers to adjustable settings determined before training a learning model or data-collection efforts for further improvement.[23]

3. Learning Curves

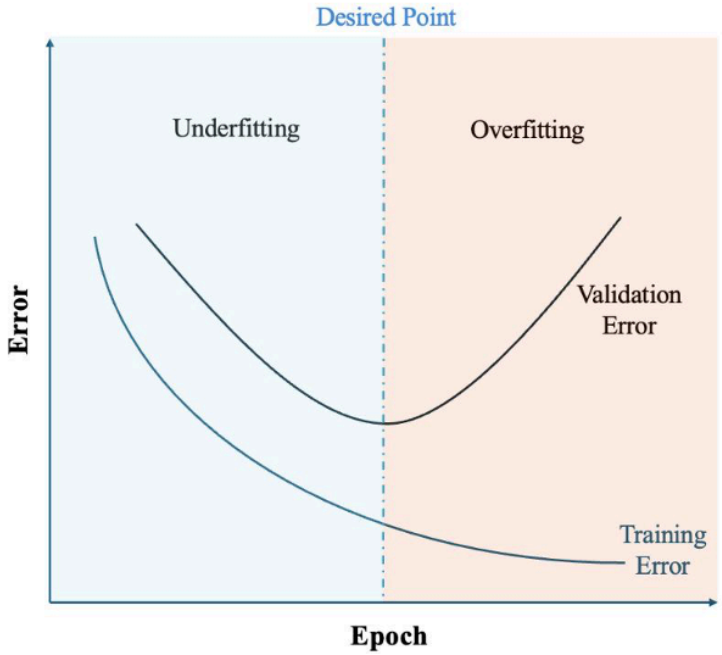

Figure 10: Learning Curve and Overfitting

Learning curves are graphs that plot a machine learning model's performance over time, measured separately on the training data and unseen validation data. They are

[21] *See* Trevor Hastie et al., The Elements of Statistical Learning: Data Mining, Inference, and Prediction 229 (2nd ed. 2009).
[22] *What is Overfitting?*, Amazon Web Services, https://aws.amazon.com/what-is/overfitting/ (last visited Aug. 8, 2024).
[23] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn*, Keras, and TensorFlow 118–20 (2nd ed. 2019).

primarily used to diagnose whether a model is underfitting or overfitting. Underfitting may be occurring if a learning curve shows that the model performs poorly on both the training and validation sets.

After adjusting and retraining, developers re-examine the learning curves. If the model now performs well on the training data but poorly on the validation data, this indicates overfitting. Learning curves guide developers through a dynamic process: first identifying underfitting, making adjustments, and then checking for signs of overfitting as complexity is introduced. To use this testing mechanism, the dataset is divided into two sets: the training set and the validation set.[24] First, the model will learn using the training set. Then, the model's performance will be tested on the validation set. The error for both sets will be logged as the training proceeds. With the epochs of the training data increasing, the error in the training set naturally decreases as the model understands the training data. The validation error will also decrease at first as the model fits the representative data patterns. This trend will continue, represented by the blue underfitting zone in Figure 10, until the desired point is reached.[25] If the model continues to study the training set, the validation error will increase because the model is overfitting the training data and collecting excessive noise. The desired point is reached before excessive training leads to overfitting. To summarize, when a model is improving its performance on the training test, and the error in the validation set starts to rise, it is a strong indicator of overfitting.[26]

II. THE ITALIAN PLUMBER PROBLEM

A. Definition of the Italian Plumber Problem

A specific issue that may arise from overfitting is the Italian Plumber Problem, illustrated in an article by Lee and Grimmelmann. Stability AI and ChatGPT were

[24] *Learning Curve To Identify Overfit & Underfit*, GeeksforGeeks, May 17, 2024, https://www.geeksforgeeks.org/learning-curve-to-identify-overfit-underfit/.
[25] Xue Ying, *An Overview of Overfitting and its Solution*s, Journal of Physics: Conf. Ser. 1168 (2019).
[26] Melanie, *Overfitting: what is it? How can I avoid it?*, Datascientest, Sept. 23, 2023, https://datascientest.com/en/overfitting-what-is-it-how-can-i-avoid-it.; Jason Brownlee, *How to Identify Overfitting Machine Learning Models in Scikit-Learn*, Machine Learning Mystery, Nov. 27, 2020, https://machinelearningmastery.com/overfitting-machine-learning-models/.

prompted to generate a picture of "Video Game Italian Plumber."[27] Each program generated an output that highly resembled Mario, the Italian plumber from the video game *Super Mario*, a copyrighted character. This issue was first brought up by Professor Matthew Sag, addressed as "The Snoopy Problem."[28] Although both ideations of the problem follow the same idea, there is a difference between how Lee, Grimmelmann, and Professor Sag approach the problem. In the "Snoopy Problem" demonstration, the AI was prompted to create an image of "Snoopy lying on a red dog house with Christmas lights on its comic."[29] In the "Snoopy Problem" demonstration, the AI's prompt explicitly invoked a known character, making clear it was a deliberate effort to reproduce protected expression. Arguably, the more troubling overfitting scenario arose when no name was mentioned and the model still generated an image of Mario's red cap, moustache, and overalls that unmistakably mirror the copyrighted character. I will examine the problem of overfitting in scenarios where even when the name "Mario" is not mentioned in a prompt, the model still produces an infringed image of the protected character.[30] Model infringement is when a character's protectable expression is reproduced so faithfully that an ordinary observer would immediately recognize it as the character's expression rather than a new work.

B. What is a Copyrightable Character?

Characters are not explicitly listed in the Copyright Act of 1967 as copyrightable subjects.[31] They are now protected under this statute due to the practical interest of promoting efficiency in infringing lawsuits. Two tests are commonly applied to determine whether a character is copyrightable.

The well-delineated test requires the court to follow a three-step evaluation.[32] First,

---

[27] Timothy B. Lee et al., *Why The New York Times might win its copyright lawsuit against OpenAI,* Understanding AI, Feb. 20, 2024, https://www.understandingai.org/p/the-ai-community-needs-to-take-copyright.

[28] Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. 295 (2023).

[29] *Id.*

[30] *Id.*

[31] 17 U.S.C. § 102(a) (2018) (listing categories of copyrightable works, such as literary works, musical works, and motion pictures, but not mentioning characters explicitly).

[32] Warner Bros. Inc. v. American Broadcasting Cos., 720 F.2d 231, 240 (2d Cir. 1983) ("A character may be copyrighted if it is sufficiently delineated to be recognizable as the same character whenever it appears, and it displays consistent, identifiable character traits and attributes").; *see also* Metro-Goldwyn-Mayer, Inc. v. American Honda Motor Co., 900 F. Supp. 1287, 1296 (C.D. Cal. 1995) (applying the

the character must possess concrete physical and conceptual attributes. Physical attributes are defined as any sensory or visual details – costume, facial features, distinctive gestures, or vocal mannerisms – that immediately identify the character.[33] Conceptual attributes include the character's inner life and context, such as backstory, motivations, moral convictions, relationships, and the environments they inhabit.[34] Second, these traits must recur consistently throughout the work, across scenes, chapters, or episodes, such that the character remains unmistakably the same "player."[35] Finally, the character must be "especially distinctive."[36] Beyond mere archetype, the combination of traits must exhibit originality or vividness, such as Sherlock Holmes's forensic catchphrases and violin-playing eccentricities, creating a character that transcends stock roles.[37] The character's expression must be so distinctive that an ordinary reader or viewer would immediately recognize it as that specific creation. For example, in *Anderson v. Stallone,* the Ninth Circuit held that the character "Rocky Balboa" was sufficiently delineated to merit protection under 17 U.S.C. § 102(a).[38] Balboa's trademark boxing attire, distinctive Philadelphia accent, and scrappy underdog ethos remained consistently expressed across multiple works. By applying the well-delineated test, the court found that these recurring physical and conceptual markers created an "original expression" separate from any single screenplay, qualifying Rocky as a copyrightable character.

The "story being told" test was applied by the Ninth Circuit in *Warner Bros. v. Columbia Broadcasting System* to determine whether a character is so central to the narrative that the story revolves around their distinctive personality, rather than merely using them as a vehicle for the plot.[39] The case discusses the extent to which a character–specifically, Sam Spade from the detective novel *The Maltese Falcon*–can receive copyright protection as an independent work, apart from the larger story. The court ultimately held that the character was not protected because he was not the story being told, but rather the agent through whom the story was told.[40] The court was

---

well-delineated character test and requiring "physical and conceptual qualities, sufficiently delineated to be recognizable, and especially distinctive").

[33] A*m. Broad. Cos.*, 720 F.2d, *supra* note 32, at 240.

[34] *Id*.

[35] *Id*.

[36] *Honda Motor Co.*, 900 F. Supp., *supra* note 32, at 1296.

[37] A*m. Broad. Cos.*, 720 F.2d, *supra* note 32, at 240.

[38] Anderson v. Stallone, 19 F.3d 62.

[39] Warner Bros. v. Columbia Broadcasting System, 216 F.2d 945 (9th Cir. 1954).

[40] *Id.*

asked whether Sam Spade, the detective protagonist of The Maltese Falcon, qualified for independent copyright protection. Under this test, a character is protectable only if it constitutes the story. If it serves merely as "a chessman in the game of telling the story,"[41] it falls outside the copyright's scope. Although Sam Spade is a vividly drawn detective, the Ninth Circuit held that the narrative's core idea lay in the quest for the stolen statuette, not in Spade's persona, rendering him a storytelling vehicle. As a result, Spade could not claim standalone protection, as only characters integral to a story's substance satisfy the "story being told" test and merit copyright coverage. This delineates the scope of protection given to a fictional character.[42]

C. Element-by-Element Analysis of Mario as a Copyrightable Character

The well-delineated test and the "story being told" test are applied here to determine whether Mario – the iconic Italian plumber created by Nintendo – qualifies for independent copyright protection.

1. Well-delineated Test

Under the well-delineated framework, a character is protectable if it (a) possesses concrete physical and conceptual attributes; (b) can be recognized "instantly and unmistakably" across different works; and (c) reflects more than stock or functional traits.[43] First, Mario's trademark red cap with a capitalized "M," blue jumpsuit, and white gloves, with a signature mustache, constitute a distinctive visual costume as its physical attributes. Amongst all titles, bravery, altruism, and optimism are recurring traits of Mario. The signature "It's-a-me!" exclamation and his Italian-accented English inflection reinforce a stable character blueprint.[44] Every Mario game, from Donkey Kong (1981) to Super Mario Odyssey (2017), retains these elements, making the plumber "instantly and unmistakably" recognizable as Mario. His combination of attire, speech, and narrative role transcends mere plot function, demonstrating the kind of unique expression that the Ninth Circuit recognized in *Anderson v. Stallone*.[45]

---

[41] *Col. Broad. Sys.,* 216 F.2d, *supra* note 39.
[42] *Id.*
[43] *Stallone*, 19 F.3d, *supra* note 38 (applying well-delineated test to "Rocky Balboa").
[44] Super Mario Bros. 2 (Nintendo Co. 1988).; Super Mario Odyssey (Nintendo Co. 2017).
[45] *Stallone*, 19 F.3d, *supra* note 38.

2. Story Being Told Test

This second test questions whether the character's traits are central to the work's narrative, and whether those traits drive a unique sequence of events rather than serving purely functional or background purposes.[46] Mario is invariably the protagonist whose decisions and actions shape each game's story arc. The entire narrative – from level design to boss battles – is built around his quest, rather than around interchangeable settings or secondary figures. Unlike a nonspecific avatar, Mario's personal history infuses each installment with character-driven stakes and emotional continuity.

Applying both the well-delineated test and the "story being told" test confirms that Mario possesses the requisite originality, consistency, and narrative centrality to qualify as a copyrightable character in his own right.

D. Using Overfitting as a Lens

Mario is a copyrightable character that is legally protected under the Copyright Act. However, an AI model might not recognize that it is infringing Mario's copyright when creating a picture of an Italian Plumber. Examining this phenomenon in the context of overfitting, I argue that the Italian Plumber Problem occurs because of the skewness and insufficiency of the model's training data. The character Mario is so famous that it has a near-monopolistic position in the animated Italian plumber image generation data processed by the model. Nonexistent data variation that fits this description results in the model saliently matching most characteristics of Mario.

Understanding that this issue derives from overfitting, the model trainer can apply a framework to detect and reduce infringement. Certain efforts should be made to avoid overfitting by developers, not only for the sake of model accuracy but also to prevent unlawful infringement of copyright owners. There are common ways to avoid overfitting, such as early stopping, pruning, regularization, ensembling, and data augmentation.[47] For example, data augmentation is when program developers alter data to obscure the significance of unique characteristics possessed by characters, an element protected by the well-delineated test. More weight might be placed on images of Western plumbers disassociated with Mario to reduce the linkage between Mario

[46] Nichols v. Universal Pictures Corp., 45 F.2d 119 (2d Cir. 1930) (establishing the story being told test criterion for character protection).

[47] *What is Overfitting?*, *supra* note 22.

and animated Italian plumbers. This may help avoid situations where AI produces an image of Mario for the user when they only want an animated Italian plumber. However, implementing this successfully requires preliminary knowledge of the model to identify infringing output. While Professor Grimmelmann and Lee suggested that both Stability AI and GPT-4 spit out infringing images of Mario under the prompt "An Italian plumber from a video game,"[48] when I tried this out, my attempt was not successful. I tried to instruct ChatGPT to generate an image of Mario with the same prompt used by Professor Grimmelmann and Lee, but it refused to do so and added a warning stating that, to "avoid copyright issues," it cannot draw such a picture. Taking it a step further, I asked the model directly for a picture of Mario, which did not work either. This implies that AI developers are currently making efforts to identify and avoid generating copyrighted work.

Despite these two failed experiments, ChatGPT offered me an alternative option of an image that would obscure key characteristics of Mario and would not amount to a copyright infringement. Still, the alternative image bears unignorable resemblances to Mario (a capital letter "M" on the same red color and style of newsboy hat, similar outfit style, and a classic Mario mustache) despite ChatGPT contending that the image does not constitute an infringing level of similarity.

III. APPROACHES TO HANDLING THE PROBLEM

A. Fair Use Analysis and Why It's Not a Good Fit

The fair use doctrine is a legal principle used to balance the protection of copyright owners with the use of copyrighted material. Considering that Generative AI is likely to infringe upon copyrightable work, the fair use doctrine could be used to protect the model creators from liability. Four elements are considered when evaluating if a case of copyright infringement constitutes fair use: "the purpose and character of the use, the nature of the copyrighted work, the amount or substantiality of the portion used, and the effect of the use on the potential market for or value of the work."[49]

In the 1994 Supreme Court Case *Campbell v. Acuff-Rose Music, Inc.*,[50] the court

[48] Timothy B. Lee & James Grimmelmann, *Why The New York Times might win its copyright lawsuit against OpenAI*, Understanding AI, Feb. 20, 2024, https://www.understandingai.org/p/the-ai-community-needs-to-take-copyright.

[49] *Fair Use*, Columbia U. Libraries, https://copyright.columbia.edu/basics/fair-use.html (last visited Apr. 30, 2025).

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

applied fair use with a focus on the first factor, determining when the "purpose and character of the use" was substantial enough to bypass copyright.[51] The case involved the rap group 2 Live Crew, which created a parody of Roy Orbison's song "Oh, Pretty Woman" without obtaining a license.[52] Acuff-Rose Music, which owned the copyright, sued for infringement. The court assessed the "transformativeness" of the work, placing significant weight on the merits of the party.[53] The outcome of the case determined that if a work is transformative, such that it supersedes the original purpose, the infringers may use this defense to avoid infringement claims. The subjective nature of this evaluation makes it difficult to determine when work is transformative enough to constitute a fair use defense. Since AI technology is relatively new, the unpredictability of a fair use defense presents a problem for start-ups that lack a professional legal team to assess copyright risk. OpenAI has the resources to cope with a copyright lawsuit should it arise, but a small company could go bankrupt from just hiring a team of lawyers. Potential lawsuits and statutory fines incurred by copyright infringement may deter small corporations from creating AI models. This exacerbates the oligopoly in the technology market by raising the entry barrier, ultimately hindering technological progress.

Infringement often occurs outside the control of the model's developer. Developers have little control over what the user asks the model to create, and which graphics, paragraphs, or snippets the model will generate. In the case of legally regulating AI, the second factor of the fair use analysis – purpose and character of the work – would be hard, if not impossible, to determine. Models are engineered to detect and reproduce patterns based on statistical correlations in their training data. Their core function is to generate content that aligns with the input they receive by drawing from patterns. This process is not guided by an understanding of legality or originality, but by probabilistic associations. As a result, any infringing output produced by AI is not the product of conscious decision-making, but rather an incidental byproduct of the design to optimize for relevance, coherence, and user satisfaction.

Another issue with applying the fair use doctrine to AI models is the subjectivity of the doctrine, given its nature as a four-factor aggregate test. Professor Matthew Sag expresses that the fair use analysis is not an instructive policy instrument; rather, there is a consensus that the fair use doctrine should be used to bring greater good.

---

[51] Rich Stim, *Measuring Fair Use: The Four Factors*, Stanford Copyright and Fair Use Center (2021), https://fairuse.stanford.edu/overview/fair-use/four-factors/.
[52] *Id.*
[53] *Id.*

Copyright law is not intended to prevent the mere act of copying, but rather to ensure the rights of creators.[54] A clear legal guideline is needed to ensure the integrity and consistency of a judgment by the courts and allow stakeholders to follow through. Evaluating the potential costs and benefits of AI models may justify a defense that strays from the court's role. The subjective nature of the fair use doctrine hampers the establishment of a straightforward rule. This is due to the doctrine's case-by-case application, which relies heavily on judicial interpretation rather than clear-cut standards. As AI technologies become more widespread, corporations and developers are facing lawsuits over copyright infringement–making ambiguity in fair use a pressing legal challenge in today's digital landscape. Considering infringement can occur simply from user input, the Federal Courts may also be overwhelmed with frivolous lawsuits. After exploring problems with applying the fair use doctrine, I will proceed with my recommendation to curtail fair use.

B. Approaches to Assist Fair Use

Following the argument that AI model developers should be held liable for infringing output, I propose several measures to determine liability and evaluate the infringing act. These approaches not only protect AI developers but also aim to better regulate the use of copyrighted work in the AI training process.

1. Good Faith to Avoid Liability

First, I argue that courts should take into account a company's good-faith efforts to prevent copyright infringement when determining liability. The Mario case shows that ChatGPT developers were successful in creating a model that refused to generate copyrighted work. However, it is unclear whether this standard is applied to every AI model. If developers are held liable for copyright infringements, the concern arises that liability could deter smaller AI developers from creating models. Developers may begin training models to identify copyrighted work to establish a valid defense against infringement claims. In *Sony Corp. of America v. Universal City Studios, Inc.*,[55] the Supreme Court recognized that a defendant's good faith efforts to prevent infringement may weigh against liability, particularly where the technology has substantial non-infringing uses. A good-faith defense would encourage developers to

[54] Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 Fordham L. Rev. 1887 (2024).
[55] Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417, 442 (1984).

implement safeguards without fear of disproportionate legal consequences.

2. Reduce the Opacity of Training Data

A second proposal addresses challenges in the evolving landscape of copyright and AI technology – the difficulty in the evidence collection process for AI infringement lawsuits. Owners of copyrighted work would be better protected if they could track where and how their work is used for AI training models. Individuals are often unaware of how their work contributes to the output of AI models, given the massive amount of data models are trained on.[56] This is obvious in the case of *Anderson v. Stability AI,*[57] where a group of artists sued Stability AI, Midjourney, and DeviantArt for allegedly using their copyrighted artworks without permission to train generative AI models like Stable Diffusion. In October 2023, the Northern District of California dismissed nearly all claims, including those of vicarious copyright infringement and violations under the DMCA. The dismissals were largely due to insufficient evidence, as plaintiffs struggled to demonstrate exactly how their works were used. This is because AI developers do not keep old training data, reducing the transparency of the process. Individual creators would also have difficulty sorting through millions of data points to find evidence of a model scraping the expression off a work. Implementing regulations that require AI developers to log their training data sources would facilitate greater accountability.

3. Promote a Model to Evaluate Overfitting Score

Another measure to ensure responsible AI development is to create a general algorithm to evaluate the overfitting score of the model developed in the context of copyright infringement. Utilizing overfitting detection techniques, such as K-fold Cross-Validation and learning curves, allows an AI developer to determine the potential infringement level of a model's output. If an AI model exhibits minimal signs of overfitting during testing, developers may reasonably interpret this as evidence that the model is not memorizing or reproducing copyrighted material. From this perspective, releasing the model to the public appears justified, as the company has taken responsible, preemptive steps to avoid infringement. This supports the argument that the developers acted in good faith – an important factor in mitigating liability.

[56] Pamela Samuelson, *Generative AI Meets Copyright*, 381 Science 158 (2023).
[57] *Andersen v. Stability AI Ltd.*, No. 3:23-cv-00201, 2023 WL 6292920 (N.D. Cal. filed Oct. 30, 2023).

However, because generative models are probabilistic and capable of producing unforeseen outputs, there remains a risk that future users could prompt the model in ways that lead to infringement. The significance here lies in the legal tension between responsible design and unpredictable use: a model's initial compliance may not guarantee continued compliance, but it may still shield developers from liability if good faith efforts can be demonstrated.

4. Assign Liability to Inappropriate Use of "Infringing" Output

A fourth proposal to address AI copyright infringement is to hold end users accountable for unlawful uses of infringing output instead of AI developers. The liability should be shifted to end users who misuse the infringing output for public or commercial purposes, given that the developers can prove they have taken measures to identify infringing output and warn users of unlawful practices.

Compensation for copyright owners over the potential loss of revenue may also be charged via taxation on AI developers. This proposal draws on principles from the Audio Home Recording Act (AHRA) of 2017,[58] which established levies on the distribution of digital audio recordings to fund compensation for copyright owners whose work had been infringed. This shielded the recorder hardware manufacturers and distributors from noncommercial use of the copyrighted work.[59] Similarly, a levy system could be incorporated into the AI market. This measure offers a policy-based approach to compensating copyright owners for potential losses by establishing a fund to subsidize creators whose works have contributed to the public benefit.

IV. FURTHER DISCUSSION: FAIR USE HAS AN OVERFITTING PROBLEM

An argument can be made that the fair use doctrine has an overfitting issue. I designed a table to demonstrate this argument by comparing overfitting in fair use cases to overfitting in data.

Overfitting in data	**Overfitting in Fair Use cases**
Relevant/ Valid Data	Facts essential to the merits/ ruling of the case
Noises	Irrelevant facts specific to the case ruling
The appropriate Pattern/	Legal Doctrine to be applied

[58] Audio Home Recording Act, 17 U.S.C. § 1001 (1992).

[59] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 45-97 (2017).

Algorithm abstracted	
Model designed by AI developers	Lower courts
Output	The ruling that the Supreme Court gave or would give had it taken the case

Under the American case law tradition, each time a lower court rules on a specific case, it predicts how the Supreme Court would have ruled. Lower court judges play a similar role to AI models: they (1) were given the input of past cases – the facts and previous rulings (2) would need to dissect the precedents to separate noise, irrelevant facts of the case that do not contribute to the merits of the ruling, from representative data that are necessary to determine the outcome of the case, and (3) extract the appropriate judicial doctrine behind the opinion to decide on the appropriate principle for the legal issue. The judicial system, like AI developers, refines the "model" by establishing the line of appeals. This system upholds or overturns the previous decision, and ultimately might require the Supreme Court to provide the final decision.

Using this parallel relationship, we can further analyze the fair use doctrine and why it has an overfitting problem, especially within AI technology. As the previous section discussed, a key factor contributing to overfitting is insufficient data. With no previous ruling specifically tailored to AI, lower courts make a ruling without precedent to consult. Additionally, noise present in the data – relevant facts that do not contribute to the merits of the case – would also be a challenge for judges. This can lead to misinterpretations, where judges may wrongly treat surface-level factual similarities as controlling. For instance, if a lower court interprets a case in a way that contradicts how the Supreme Court would rule, it may be because it mistakenly identifies insignificant facts as central to a precedent. The unexplored legal vacuum created by the emergence of new technologies may experience a bigger overfitting issue. This is demonstrated by the court's difficulty using the fair use analysis concerning copyright issues within AI models. In the future, it would be interesting to view these problems from a data analysis perspective and consider solutions for detecting and reducing the problem of overfitting.

CONCLUSION

AI has demonstrated its ability to be comparable to the human creator. This new reality calls for the courts to address the legal gaps left by these technological advancements. The intersection of AI and law is a complex and evolving field. This paper has focused on a specific aspect of AI's potential infringement: overfitting during model training. By dissecting this phenomenon and exploring its root causes, this paper strives to bridge the gap between technical intricacies and legal considerations as a detailed reference to legal practitioners. The paper centers on the Italian Plumber Problem and provides updated efforts made by AI developers such as ChatGPT. Alongside an evaluation of the fair use analysis to highlight the unpredictability and limitations of existing frameworks, I offered supplementary approaches that may better address AI's copyright infringement issues. Finally, I pointed out a surprising yet intriguing parallel between model development and judicial rulings that underscores a critical insight; just as overfitting is a concern in AI, the same principles apply to the limitations of fair use in the context of generative AI technology. We must continue to refine our legal frameworks, ensuring that they are robust enough to handle the challenges posed by revolutionary technology.