INFORMATION SYSTEMS EDUCATION JOURNAL

Volume 23, No. 5
September 2025
ISSN: 1545-679X

In this issue:

4. ***Invited Paper***
Treating problematic smartphone usage in the TikTok era: cultural dimension bias in the current research
David Firth, University of Montana
Tyler Johnson, University of Montana
Shawn F. Clouse, University of Montana
Patricia Akello, University of Montana

20. **Blockchain Education: Evaluating Programs, Curricula, and Integration with Emerging Technologies in Higher Education**
Tan Gürpinar, Quinnipiac University
Aashman Verma, Quinnipiac University
Lori N. K. Leonard, The University of Tulsa
Kiku Jones, Quinnipiac University
Wendy Ceccucci, Quinnipiac University

30. **Exploring VR-Enhanced Learning in Business Education: A Multi-Site Study**
Kevin Mentzer, Nichols College of Business
Mark Frydenberg, Bentley University
Suhong Li, Bryant University

44. **Assessing the Impact of Prerequisite Courses on Student Performance in Database Normalization**
Kevin Slonka, Saint Francis University
Matthew North, Utah Valley University
Neelima Bhatnagar, University of Pittsburgh
Anthony Serapiglia, Saint Vincent College

53. ***Teaching Case***
Countering the "Plagiarism Slot Machine": Protecting Creators and Businesses from AI Copyright Infringement
Christine Ladwig, Southeast Missouri State University
Dana Schwieger, Southeast Missouri State University
Reshmi Mitra, Southeast Missouri State University

iscap
Information Systems & Computing Academic Professionals

The **Information Systems Education Journal** (ISEDJ) is a double-blind peer-reviewed academic journal published by **ISCAP** (Information Systems and Computing Academic Professionals). Publishing frequency is five times per year. The first year of publication was 2003.

ISEDJ is published online (https://isedj.org). Our sister publication, the Proceedings of the ISCAP Conference (https://iscap.us/proceedings) features all papers, abstracts, panels, workshops, and presentations from the conference.

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the ISCAP conference. All papers, whether award-winners or not, are invited to resubmit for journal consideration after applying feedback from the Conference presentation. Award winning papers are assured of a publication slot; however, all re-submitted papers including award winners are subjected to a second round of three blind peer reviews to improve quality and make final accept/reject decisions. Those papers that are deemed of sufficient quality are accepted for publication in the ISEDJ journal. Currently the target acceptance rate for the journal is under 35%.

Information Systems Education Journal is pleased to be listed in the Cabell's Directory of Publishing Opportunities in Educational Technology and Library Science, in both the electronic and printed editions. Questions should be addressed to the editor at editor@isedj.org or the publisher at publisher@isedj.org. Special thanks to volunteer members of ISCAP who perform the editorial and review processes for ISEDJ.

2025 ISCAP Board of Directors

Amy Connolly
James Madison University
President

Michael Smith
Georgia Institute of Technology
Vice President

Jeff Cummings
Univ of NC Wilmington
Past President

David Firth
University of Montana
Director

Mark Frydenberg
Bentley University
Director/Secretary

David Gomillion
Texas A&M University
Director

Leigh Mutchler
James Madison University
Director

RJ Podeschi
Millikin University
Director/Treasurer

Jeffry Babb
West Texas A&M University
Director/Curricular Matters

Eric Breimer
Siena College
Director/2024 Conf Chair

Tom Janicki
Univ of NC Wilmington
Director/Meeting Planner

Xihui "Paul" Zhang
University of North Alabama
Director/JISE Editor

Copyright © 2025 by Information Systems and Computing Academic Professionals (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to Paul Witman, Editor, editor@isedj.org.

INFORMATION SYSTEMS EDUCATION JOURNAL

Editors

Paul Witman
Editor
California Lutheran University

Thomas Janicki
Publisher
U of North Carolina Wilmington

Dana Schwieger
Associate Editor
Southeast Missouri State University

Kevin Mentzer
Associate Editor
Nichols College

Ira Goldstein
Teaching Cases & Exercises Co-Editor
Siena College

Michelle Louch
Teaching Cases & Exercises Co-Editor
Duquesne University

Donald Colton
Emeritus Editor
Brigham Young University Hawaii

Jeffry Babb
Emeritus Editor
West Texas A&M University

Teaching Case

Countering the "Plagiarism Slot Machine": Protecting Creators and Businesses from AI Copyright Infringement

Christine Ladwig
cladwig@semo.edu

Dana Schwieger
dschwieger@semo.edu

Reshmi Mitra
rmitra@semo.edu

Southeast Missouri State University
Cape Girardeau, MO 63701

Hook

In 2022, Artist Kelly McKernan began seeing her dreamy, sci-fi style of creative images online, but there was a big problem—it wasn't her work. McKernan discovered that people around the world were using the prompt "in the style of Kelly McKernan" and generating images through AI platforms such as Midjourney and Stable Diffusion. So, in January 2023, McKernan joined artists Sarah Andersen and Karla Ortiz in filing a copyright infringement lawsuit against these AI generative program developers claiming that the companies infringed the rights of millions of artists. The complaint alleges that these AI platforms scraped five billion images from the web during the training of their systems and are using these copyrighted materials without the consent of the original artists. How can computer scientists and programmers contribute to finding a solution to this increasingly complex problem?

Abstract

The rapid rise of AI use is creating some very serious legal and ethical issues such as bias, discrimination, inequity, privacy violations, and—as creators everywhere fear—theft of protected intellectual property. Because AI platforms "learn" by scraping training materials available online or what is provided to them through their human programmers, these systems can easily capture copyrighted expressions, such as song lyrics, computer code, stories, or images, and use them to generate new works without attribution. This rise in AI use of protected material is spawning an array of legal actions as artists, programmers, writers, photographers and other creative individuals witness the erosion of their value in the marketplace and the world. As students prepare to enter the field, they need to be aware of legal issues and concerns that they may face and methods for addressing them. This case focuses on the problem of AI copyright infringement of art and includes an exploratory exercise that introduces students to the act of "scraping"—a primary AI training method by which copyrighted works may be vulnerable to potential infringement.

Keywords: Teaching Case, Screen Scraping, Data Pools, Metadata, AI Copyright Infringement, AI Training

Recommended Citation: Ladwig, C., Schwieger, D., Mitra, R., (2025). Countering the "Plagiarism Slot Machine": Protecting Creators and Businesses from AI Copyright Infringement. *Information Systems Education Journal*, v23, n5 pp 53-61. DOI: https://doi.org/10.62273/PCXH9792

Countering the "Plagiarism Slot Machine": Protecting Creators and Businesses from AI Copyright Infringement

Christine Ladwig, Dana Schwieger and Reshmi Mitra

1. INTRODUCTION

The rapid advancements of generative artificial intelligence (AI) are impacting almost every aspect of life… even the art world. Creative artists are finding that they are competing not only with other artists for commissions, but also with generative AI programs trained on versions of their own work. Issues concerning generative AI training datasets continue to surface. Researchers in the Data Provenance Initiative evaluated 1,858 of the most popular open-source publicly available datasets used to train natural language processing models (e.g., GitHub, Hugging Face, Papers With Code, etc.). Through their audit, they found that over 70 percent of the datasets had no data licenses. For those that did have licenses, "roughly half were incorrect" and "29 percent of the incorrect licenses were more permissive than the dataset creators had intended" (Gent, 2023).

Copyright infringement lawsuits against tech companies are being filed and as organizations, such as *The Atlantic* magazine, make searchable databases of AI dataset resources available to artists, will continue to increase (Gent, 2023). Legislation for protecting creative works is in the process of being developed; however, this is a relatively new territory.

With these advancements in mind, students preparing for careers in industry need to be aware of and develop skills for addressing dataset concerns. This case introduces students to copyright concepts and issues associated with machine learning training datasets through the examination of current legal actions, responses by industry standard setting bodies to putting safeguards in place, and a hands-on exploratory exercise in which students scrape a site for images and examine their metadata.

2. PLAGIARISM SLOT MACHINE

Kelly McKernan is a very successful fantasy artist with famous clients including Horse Comics, Stranger Things, and ImagineFX. Within the last few years, another group of artists have been posing a challenge to McKernan's position in an already limited marketplace—AI generative programs such as Midjourney and Stable Diffusion. It was discovered that people around the world were using the prompt "in the style of Kelly McKernan" and generating images through these AI platforms. Over 11,000 images in McKernan's distinctive style were found on one platform server alone, and all were generated without the artist's consent or input (Chow, 2023).

Artists' Style Reproduced by Generative AI

McKernan noted that opportunities for creating art are disappearing as AI generative programs increase in use; for instance, prior to the rise in Midjourney/Stable Diffusion, McKernan was securing multiple commissions per month. But now, some of those opportunities are flowing to the AI generative programs: "It's…pretty wild to know that instead of hiring me (McKernan) for a book cover, someone can just go into a program, use my name to emulate something close enough and good enough, at a fraction of the price" (Chow, 2023). And the feeling of violation by artists is significant, according to McKernan "AI is not a tool, it's a plagiarism slot machine" (Dean, 2023).

McKernan and Other Artists File Lawsuit

In January 2023, Kelly McKernan joined artists Sarah Andersen and Karla Ortiz in filing a copyright infringement lawsuit against these AI generative programs claiming that companies such as Midjourney and Stable Diffusion infringed the rights of millions of artists. The complaint alleges that these AI platforms scraped five billion images from the web during the training of their systems and are using these copyrighted materials without the consent of the original artists. The suit also stated that, "These companies benefit commercially and profit richly from the use of copyrighted images…the harm to artists is not hypothetical, as generative AI art is already sold on the internet, siphoning commissions from the artists themselves." The original suit was filed in January 2023 and amended in November of the same year to add seven plaintiffs and one new defendant, Runway AI.

Tech Companies Sued

Artists are not the only creators that are feeling the sting of AI's dominance in their industry. On November 3, 2022, Microsoft Corporation and its computer code-sharing website GitHub, as well as artificial intelligence firm OpenAI, were sued in the U.S. District Court for the Northern District of California. The class-action complaint (J. DOE 1, et al., Plaintiffs, v. GITHUB, INC., et al., Defendants) claimed that the companies' AI-powered programming tool Copilot infringed copyright by using millions of lines of human-written code without proper attribution. According to reporters for NewScientist, this is the "first big copyright lawsuit over AI and potential damages could exceed $9 billion" (Wilkins, 2022).

Software Programmers' Work Violated

Around the same time that the action against Microsoft and Open AI was filed, internet chatter began to blanket the Web with similar claims of AI infringement against software coders, artists, writers, and other content creators. For instance, in October 2022, Texas A&M University Computer scientist Tim Davis claimed on Twitter (now X) that the Microsoft-owned AI programming assistant Copilot "emits large chunks of my copyrighted code, with no attribution, no LGPL license." (Davis, 2022) The "LGPL" Davis mentions is a type of Open-Source use permission—the Lesser General Public License—which makes the code available for use to anyone if they adhere to the license requirements, such as attribution (which identifies the copyright holder of the work being reused—in this case developer Davis). Davis laments the fact that not only is his copyright being reaped and infringed by commercial AI generative programs, but the copied code is also enclosed behind a paywall, defeating his intent to make it available for free (with attribution).

New York Times Files Lawsuit

In addition to artists and coders, there have been reports of widespread AI infringement of other content publishers including newspapers and journals. For example, at the end of 2023, The New York Times (*The Times*) joined the copyright infringement fray, also with an action against Microsoft and OpenAI, alleging that the Large Language Machines (LLMs) were copying and using millions of the newspaper's copyrighted articles.

The Times stated in its complaint that "Through Microsoft's Bing Chat (recently rebranded as "Copilot") and OpenAI's ChatGPT, Defendants seek to free ride on The Times' massive investment in its journalism by using it to build substitutive products without permission or payment." (The New York Times Company v. Microsoft Corporation (1:23-cv-11195), 2023). The complaint describes the infringing LLMs as containing copies of article content, which is then generated in prompt responses as text that "is verbatim, closely summarizes it [The Times' content], and mimics its expressive style." The Times submitted thousands of pages of exhibits demonstrating the alleged infringement, with entries dating back to the 1950s. The Times also pointed out the benefits of the AI developers' unauthorized use: a trillion dollar increase in Microsoft's market capitalization due to Copilot, and a $90 billion valuation for OpenAI's ChatGPT. (The New York Times Company v. Microsoft Corporation (1:23-cv-11195), 2023).

Clearly, AI generative programs are creating havoc through their training and operational actions, imperiling the livelihoods of creators all over the world. An engineer who spoke to author James Vincent of the *Verge* about AI copyright infringement shared this sobering sentiment: "If you take my attribution off, my career is over, and I can't support my family, I can't live." (Vincent, 2022).

Copyright vs. Trademark vs. Patent

While protecting Kelly McKernan's artwork and the stories written by New York Times' journalists is best accomplished by copyright, the latter is not the only form by which creators may safeguard their intellectual property assets. For example, inventions are typically protected through the use of the "patent", which is a grant from a government authority made in exchange for public disclosure. U.S. issued patent protection allows the inventor (or patent owner) to exclude others from using, making, or selling the invention for a limited time—usually around 20 years from the filing date of the patent application. Patent protection is most valuable for those creations that can easily be reverse engineered, as it allows the inventor to initially license or practice their invention without competition (ideally). Trademark is another protective measure intended for brand names, logos, slogans, and other identifiers that distinguish goods or services in the marketplace. The U.S. trademark can last indefinitely as long as it is renewed. There are intellectual property assets which have these multiple forms of protection attached—the smartphone, for instance. The software (operating system) and interface design of most phones are protected by copyright; the brand name, logo and design protected by trademark; and hardware components and touchscreen technology are

typically utility inventions safeguarded through issued patent(s). However, when considering the rights associated with creative works of art, the rules associated with copyright must be followed.

3. THE NATURE OF COPYRIGHT AND THE NATURE OF INFRINGERS

The U.S. Copyright Office cites 1790 as the year that Article I, Section 8 of the United States Constitution codified the belief that "authors of a work may reap the fruits of his or her intellectual creativity" through federal protection. Although limited in scope when initially approved by Congress—protecting only books, maps, and charts for a 14-year period—the Copyright Act of 1790 has evolved to modernly safeguard a wide breadth of original works of authorship, including "literary, dramatic, musical, architectural, cartographic, choreographic, pantomimic, pictorial, graphic, sculptural, and audiovisual creations." (U.S. Copyright Office, n.d.).

Copyright Law

Copyright law is intended to protect the works of "human" creators by granting the right to: "(1) reproduce the copyrighted work; (2) prepare derivative works based on the original; (3) distribute copies of the work; (4) perform the copyrighted work publicly; and (5) display the copyrighted work publicly" (Tyser, 2024). A work created solely by AI, according to the U.S. Copyright Office cannot be copyrighted (US Copyright Office, 2023). In AI infringement cases, the AI is trained by making a copy of the original work, possibly violating the creators' right to reproduce. Plaintiffs in these lawsuits have also alleged that any AI model is infringing due to containing compressed copies of the originals, and the generation of "new" images or works is a violation of the right to prepare derivative works.

Copyright Benefits

In granting a period through which artists, musicians, writers, and other creators can protect their works from theft, copyright laws provide not only the benefit of economic compensation for the copyright holder, but also rewards for the public as the works are created and disseminated. There is evidence that these benefits are significant for creators and distributing entities. In 2022, the International Intellectual Property Alliance (IIPA), an organization representing U.S. Copyright-based industries, reported that copyright-dependent companies added $1.8 trillion to the U.S. GDP (7.76% of the entire U.S. economy) and employed 9.6 million American workers in 2021 (Stoner & Dutra, 2022).

OpenAI's Perspective

Copyright protection is widely considered to be an appropriate and just reward for creators, providing attribution and compensation for their efforts. However, not all parties are aligning with the concept of compensating individuals such as artist Kelly McKernan for the value of their efforts. AI generative programs are indiscriminately scraping the internet and making copies of, and consuming, billions of copyrighted images, writings, and code without attribution or compensation to the creators. For instance, responding to a 2019 "Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation from the United States Patent and Trademark Office ("USPTO")", OpenAI, LP—inventor of the well-known AI generator ChatGPT argued that in order to "benefit all of humanity" and be "compelling to humans," machine learning must "analyze large corpora (which necessarily involves first making copies of the data to be analyzed)" by accessing millions of documents, photographs, images, text, and audio. The company suggests that the "use of entire works is reasonably necessary" to creating AI systems, and that because the training is done by a machine, and not a human, there is no lessening of the market or value of copyrighted works. (Chadwick, 2024).

ChatGPT also responded in December 2023, to an inquiry into LLMs by the U.K. House of Lords Communications and Digital Select Committee and stated: "Because copyright today covers virtually every sort of human expression–including blog posts, photographs, forum posts, scraps of software code, and government documents–it would be impossible to train today's leading AI models without using copyrighted materials." (HOL CDSC, 2023). The company also added that "Limiting training data to public domain books and drawings created more than a century ago might yield an interesting experiment but would not provide AI systems that meet the needs of today's citizens." (Chadwick, 2024).

The Copyright Alliance's Letter to Congress

In a December 2023 letter to the U.S. House of Representatives Subcommittee on Courts, Intellectual Property, and the Internet, Copyright Alliance (a not-for-profit copyright holder advocate) CEO Keith Kupferschmid emphasized that copyright infringement by AI platforms was significant and alarming, having "increased exponentially and grown in sophistication, causing widespread harm to the economic and creative vibrancy of the copyright community." (Kupferschmid, 2023). Kupferschmid explains that "A number of recently filed lawsuits allege

that leading AI developers use datasets to train their AI models that contain unauthorized versions of hundreds of thousands of literary works, many of which are scraped from notorious "shadow library" piracy websites. While these and similar large-scale rogue websites have harmed copyright owners for years, the problem is compounded by AI developers that scrape and ingest the stolen copyrighted materials to "train" generative AI models" (Kupferschmid, 2023). Therefore, in addition to indiscriminately scraping the internet for easily accessible copyrighted materials, Kupferschmid suggests that AI platforms are also training on protected works from piratical websites. It is clear from the associations advocating for copyright holders, as well as the entities that benefit from copyrighted creations, that AI infringement is a serious issue that requires an expedient resolution.

4. AN IMMEDIATE SOLUTION

On May 8, 2024, California Federal Judge William Orrick gave a green light to creator Kelly McKernan's case against the AI generator companies. Orrick stated that the now *ten* artists (the case began with three) behind the lawsuit "had plausibly argued that Stability, Midjourney, DeviantArt, and Runway AI copied and stored their work on company servers and could be liable for using it without permission." (Brittan, 2024) Yet, as the case moves forward, the artists are still experiencing infringement as AI generative programs copy and incorporate digital images of their artwork and offer it to their own "creators/subscribers" protected behind a paywall. Users (including businesses) of generative AI are also nervous about the potential liability if creators are able to demonstrate infringement. Successful copyright cases have awarded billions of dollars in damages. The urgency of the situation is therefore dire as these lawsuits proceed through the courts and requires an immediate solution.

5. PROTECTION FOR CREATORS AND GENERATIVE AI DEVELOPERS

Under the just and necessary presumption that creators such as artist Kelly McKernan are entitled to copyright protection, as well as compensation and/or attribution for their creations, what then is the solution to protecting rights holders while concurrently creating a path for AI developers to legally use their works for AI training purposes? In order to protect these artists and provide some measure of control, as well as compensation and/or attribution, computer scientists and programmers are called upon to design generative AI training pools with a conscionable effort to respect the rights of the product creators. The question becomes, how is this done? The first step in avoiding copyright infringement is for programmers and image data pool designers to become familiar with classifying and identifying graphic image files through metadata.

6. IMAGE METADATA FAMILIARITY

There are many copyright infringement lawsuits currently working their way through the court system, and without a definitive ruling on whether AI generative programs are infringing creators' copyright, it is prudent for potential users of internet content to ensure that they are using materials that are copyright free. All graphic image files contain metadata, which is embedded information such as technical data, descriptive information, and copyright details. The International Press Telecommunication Council (IPTC) Photo Metadata specifications have been in use since 1995. Google has been using IPTC photo metadata fields since Autumn 2018 (IPTC, 2024c).

Metadata properties are grouped into Administrative, Descriptive and Rights-related properties. In response to AI data mining concerns raised by image rights owners, the IPTC, in close collaboration with the Picture Licensing Universal Systems (PLUS) Coalition, released a new version of their IPTC Photo Metadata Standard containing two new properties, Data Mining and Other Constraints on October 4th, 2023 (IPTC 2024b). Later that same month, IPTC announced that similar capabilities were added to their IPTC Video Metadata Hub version 1.5 (IPTC, 2024a).

Photo Metadata Properties

The IPTC Data Mining property contains a standard list of values that creators can use to indicate if the image can be used for AI or Machine Learning purposes. The prohibition can be set to "Prohibited except for search engine indexing" to "only permit[s] data mining by search engines to the public to identify the URL for an asset and its associated data (for the purpose of assisting the public in navigating to the URL for the asset), and prohibits all other uses, such as AI/ML training." (IPTC, 2024b) The User Note associated with the Data Mining property warns, "Similarly, the absence of a prohibition does not indicate that the asset owner grants permission for data mining or any other use of an asset." (IPTC, 2024b)

The IPTC Other Constraints property is a text-based property that was also added allowing creators to include finer detail in a human readable format (Quinn, 2023). Upon their release, the IPTC encouraged developers of graphic and video software tools, as well as generative AI crawling engines, to incorporate the new properties into their software (IPTC, 2024a). Thus, digital media creators can use these field properties to identify if their work is copyrighted and requires attribution or if the work can be used in a training data set. These properties can then be utilized by training pool creators to identify the type of attribution the media creator wishes to receive and whether or not to include the file in the pool.

Thus, with the ability to specifically associate the intentions of the creator with their works, programmers will need to incorporate processes and coding into their programs to abide by those intentions. Failure to incorporate such consideration may cause their works to infringe upon a work's copyright and place their employer at risk for possible lawsuit(s).

7. CONCLUSION

If the courts determine that the "copies" AI generative programs make during training are infringing copyright, generative AI program developers will need to be able to access training data pools of valid, usable data. Additionally, graphic data used by businesses should be copyright-free or licensed; users are not protected from copyright violations even if they are using a third-party AI generative program (both are considered to be infringing copyright). Therefore, as computer science students prepare to embark on careers that will most likely involve working with artificial intelligence, it is imperative that they are aware of, and consider, content creator rights and intentions for their works. As AI capabilities are being embedded in an increasing number of programs, faculty will need to broaden the scope and coverage of related content. To illustrate the underlying thoughts, concepts, and processes that should be considered when developing a generative AI training pool, the next two sections provide questions for discussion and an exercise that introduces students to the methods by which AI "scrapes" the internet to appropriate images.

8. QUESTIONS FOR DISCUSSION

This section can be used for case and concept discussion. After reading the case, answer the following questions.

1. What problems were artists, programmers, and content publishers (e.g., newspapers and journals) experiencing?
2. What is a copyright? Why is copyright—and not patent or trademark protection—the best protection for artwork?
3. What is the purpose of a copyright?
4. What does a copyright grant to the copyright holder?
5. What court cases were discussed in the case?
6. What logic have organizations such as OpenAI and Microsoft offered to support their use of copyrighted works?
7. What is image file metadata?
8. What metadata data mining properties are available?
9. What are some of the risks associated with training AI by scraping the internet?
10. One proposal to protect copyrighted works is to create dedicated training pools for AI. What do you think about this option?
11. Using a generative AI tool (ChatGPT, Microsoft Copilot, Google Gemini, etc.), find current court cases in which generative AI programs have infringed upon copyrights. Summarize what you find.
12. Using a generative AI tool, research how a generative AI tool can avoid copyright infringement when using training data. Summarize what you find.
13. Using a generative AI tool, what is the current status of Kelly McKernan's lawsuit?

9. DATASET DEVELOPMENT EXERCISE

A dataset development exercise with discussion questions can be found in the appendix. In this exercise, students will use a web scraping tool to garner images and image metadata from the internet based upon a selected theme. Once the images are retrieved, the data set will be filtered and cleaned to provide a usable data pool. Students are then asked to conduct an analysis of the pool images to identify ownership and copyright issues and then write a report of their findings. Discussion questions are provided that faculty may wish to discuss in class addressing the different types of screen scaping tools used, striking a balance between quantity and quality in light of copyright concerns, and ethical and legal considerations that must be considered as the pools are developed.

The assignment was not intended as a worksheet exercise but rather was created as an exploratory assignment for students to research and determine the best choices for fulfilling the needs

of program development for the assignment. Because AI "scrapes" the internet for images like the copyrighted artwork at issue in the case, the authors prepared this exercise to familiarize students with these processes as well as the limitations and benefits of scraping tools.

The process of developing a data set generator for an AI tool is a significant undertaking requiring multiple phases, iterations, and development experience. The exercise described below was created for a junior level computer science data analytics or machine learning course.

The exercise was implemented in fall 2024. The course instructor focused on the mechanics of writing an AI program to perform the web scraping process rather than the copyright data. In the fall 2024 exercise implementation, the students learned how to scrape pictures from the Web, read the picture metadata, and then used that data to store the pictures in a usable format with similar dimensions. The data sets were then used to train an AI program the students had written to detect fake photos from real ones.

The students presented their results in front of the Computer Science Department's external advisory board. The board members were very complementary of the students' project and the objectives of the exercise. The students gained a significant amount of knowledge from the exercise. Future projects may also take the copyrights metadata into consideration.

10. REFERENCES

Brittan, B. (2024). Stability AI, Midjourney should face artists' copyright case, judge says. *Reuters,* Retrieved on June 12, 2024 from https://www.reuters.com/legal/litigation/stability-ai-midjourney-should-face-artists-copyright-case-judge-says-2024-05-08/

Chadwick, L. (2024). OpenAI says it's 'impossible' to train AI without copyrighted materials. *Euronews*. Retrieved June 12, 2024 from https://www.euronews.com/next/2024/01/09/openai-says-its-impossible-to-train-ai-without-copyrighted-materials

Chow, A. R. (2023, Sept. 7). Kelly McKernan, Artist. *Time*, Retrieved May 22, 2024 from https://time.com/collection/time100-ai/6309445/kelly-mckernan/

Davis, T. (2022, Oct. 15, 8:47 PM), (@DocSparse, Twitter (X)). Retrieved June 12, 2024 from https://x.com/DocSparse/status/1581461734665367554?lang=en.

Dean, I. (2023). Illustrator Kelly McKernan reveals the raw impact of AI on artists' lives. *Creative Blog,* Retrieved May 20, 2024 from https://www.creativebloq.com/features/ai-art-the-impact-of-generative-AI.

Gent, E. (2023). Public AI training datasets are rife with licensing errors: An audit of popular datasets suggests developers face legal and ethical risks. *IEEE Spectrum*, Retrieved August 12, 2024 from https://spectrum.ieee.org/data-ai

HOL CDSC (2023, Dec. 5). House of Lords Communications and Digital Select Committee inquiry: Large language models, OpenAI—written evidence (LLM0113) https://committees.parliament.uk/writtenevidence /126981/pdf/.

IPTC (2024a). *IPTC is the global standards body of the news media*. IPTC. https://www.iptc.org/about-iptc/

IPTC (2024b). *Photo Metadata Standard 2023.2*, IPTC. https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata

IPTC (2024c). *Quick guide to IPTC photo metadata on Google Images.* IPTC https://iptc.org/standards/photo-metadata/quick-guide-to-iptc-photo-metadata-and-google-images/

J. DOE 1, et al., Plaintiffs, v. GITHUB, INC., et al., Defendants, *Doe v. GitHub, Inc.*, 22-cv-06823-JST, (N.D. Cal. Jan. 3, 2024) Retrieved on June 12, 2024 from https://casetext.com/case/doe-v-github-inc-1

Kupferschmid, K. (2023, Dec. 13). Letter to U.S. House of Representatives Subcommittee on Courts, Intellectual Property, and the Internet Retrieved on May 22, 2024 from https://copyrightalliance.org/wp-content/uploads/2023/12/Copyright-Alliance-Piracy-Letter.pdf

Quinn, B. (2023). *Rights holders can exclude images from generative AI with IPTC Photo Metadata Standard 2023.1,* IPTC, https://www.iptc.org/news/exclude-

images-from-generative-ai-iptc-photo-metadata-standard-2023-1/

Stoner, R. and Dutra, J. (2022). The International Intellectual Property Alliance (IIPA): Copyright Industries in the U.S. Economy, *2022 Report.* Retrieved May 20, 2024 from https://www.iipa.org/files/uploads/2022/12/IIPA-Report-2022_Interactive_12-12-2022-1.pdf

Tyser, D. A. (2024). Rights granted under copyright law. *Bitlaw*. Retrieved December 28, 2024 from https://www.bitlaw.com/copyright/scope.html

U.S. Copyright Office (n.d.). *A brief history of copyright in the United States.* Timeline, U.S. Copyright Office. Retrieved on May 22, 2024 from https://www.copyright.gov/timeline/

U.S. Copyright Office (2023). *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*. Copyright Office. Retrieved on August 10, 2024 from https://www.copyright.gov/ai/ai_policy_guidance.pdf

U.S. District Court, S.D. New York (2023, Dec. 27). The New York Times Company v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, Open AI GP, LLC, OpenAI, LLC, OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC. (1:23-cv-11195) https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

Vincent, J. (2022, Nov. 8). The lawsuit against Microsoft, GitHub and OpenAI that could change the rules of AI copyright. *The Verge*. https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

Wilkins, A. (2022, Nov. 8). Microsoft's Copilot code tool faces the first big AI copyright lawsuit. *New Scientist*, Retrieved on June 12, 2024 https://www.newscientist.com/article/2346217-microsofts-copilot-code-tool-faces-the-first-big-ai-copyright-lawsuit/

APPENDIX – Dataset Exercise
Web Scraping for Digital Images and Image Metadata-

Objective: To use a web scraping script to collect images and their metadata according to a specified theme.

Context: The purpose of this assignment is to provide an exploratory exercise in which students learn about web scraping programs and the data that they can collect. Students will be responsible for web scraping images based on a specified theme, such as an artist's name, color palette, or genre. Image metadata will be collected and analyzed to determine what metadata is available for examination in regard to copyright issues.

Assignment Details:

1. Theme Specification:
   - Instruction: Each group will be assigned a specific theme for your web scraping task. Themes can include:
     - A particular artist's name (e.g., Vincent van Gogh)
     - A specific color palette (e.g., pastel colors)
     - A genre or style (e.g., surrealism)
   - Deliverable: Confirm the assigned theme with the instructor and document the theme in your report.
2. Web Scraping Task:
   - Instruction: Utilize web scraping tools (such as BeautifulSoup, Scrapy, or ParseHub) to collect a dataset of digital images related to the assigned theme.
     - Develop a web scraping script to automate the extraction of images and metadata. Pay attention to the website's structure and HTML elements.
     - How could the code be modified to handle potential issues such as CAPTCHA, pagination, and dynamic content loading?
   - Deliverable: A working web scraping script, a brief description of the scraping process, and suggestions for modifying the code.
3. Data Collection Requirements:
   - Instruction: Collect at least 100 digital images relevant to the theme. For each image, gather metadata including:
     - Source URL
     - Image resolution
     - File size
     - Any available licensing information
     - Any additional metadata
     - Store this data systematically for further analysis.
   - Deliverable: A dataset (in CSV or JSON format) containing the collected images and associated metadata.
4. Discussion Questions:
   - Trade-offs in Web Scraping Techniques: What are the trade-offs between using different web scraping tools? Discuss factors like ease of use, efficiency, ability to handle complex or dynamic web pages, and overall effectiveness in collecting the required data.
   - Data Collection Process: During the data collection process, what challenges did you face? Explain how you addressed those issues?
5. Deliverables:
   - A report (3-5 pages) detailing the web scraping process, data collection, issue resolution, and overview of metadata collected.
   - Content: The report should cover the following sections:
     - Introduction and objective
     - Theme specification
     - Web scraping methodology
     - Data collection summary
     - Discussion questions
   - Format: The report should be clear, concise, and professionally formatted.
   - The collected dataset of digital images in a compressed folder (e.g., zip).
   - A CSV or JSON file containing the metadata of the downloaded images.
   - Python scripts or Jupyter Notebooks used for web scraping and data analysis, with appropriate comments and documentation.