

## Artificial Intelligence, Copyright, and the Doctrine of Fair Use



Enid White  
Wyoming Department of Transportation  
Copyright ©2024. All rights reserved. Enid White

How AI affects the Doctrine of Fair Use is one of the hottest topics out there. On one side you have authors, writers, researcher, Liberians, and on the other side you have AI designers, developers, and deployers. One side wants to guard against infringement and keep the doctrine of fair use intact. The other says because of the doctrine, there is no infringement of anyone's rights. This has led to numerous lawsuits and the court dockets are becoming backlogged. Authors are suing publishers, and AI companies enmass. Copyright infringement claims, and cease and desist letters are also on the rise. Most of the suits have elements which revolve around the Doctrine of Fair Use. The Executive Branch is also going to plan a pivotal role in how future laws on this matter evolve. To understand the doctrine and how it is affected by AI I want to start with just some basics.



## Legal Definition of Artificial Intelligence

15 USC 9401(3) defines AI as:

**... a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments.** Artificial intelligence systems use machine and human-based inputs to— (A) perceive real and virtual environments; (B) abstract such perceptions into models through analysis in an automated manner; and (C) use model inference to formulate options for information or action.

1

During the summer RAC meeting there were a lot of questions on how AI should be defined. The Federal Government defines it in 15 USC 9401(3). The key to all definitions, whether you use the federal definition, a state definition, or even a corporate definition, is the blending of machine and human interaction.



## Pros of using AI in Research Documents

- Easier to summarize and translate your work.
- Helps with grammar, structure, citations.
- AI cuts down the time to perform a task.
- AI enables the execution of complex tasks.
- Streamlines work loads.
- Automates repetitive tasks.
- Cost reduction.
- Easier data acquisition and analysis.
- Streamlines literature review.

2

DOT research centers are finding that the use of AI tools when writing reports and implementing research projects is a double-edged sword. The use of the tools may help the authors check grammar and format sentence structure, insert alternative text, and it may allow authors to perform literature reviews easier, and it just generally makes it easier to complete your work.




## Cons of using AI in Research Reports

- Unknown Plagiarism and Copyright infringement.
- Quality assurance issues.
- Flawed references.
- Invalidity of information.
- False logic/reverse curse.
- Hallucinations/fabricated knowledge.
- Poisoned data.
- Greater access from predatory journals.
- Fake papers and systematic manipulation of peer review articles.
- Bias.
- Infringement on privacy, confidential information, trade secrets.
- Lack of creativity and emotion.
- Mimics others work and writing styles.
- May infringe on consumer protection laws.
- Hard to opt out of the AI scraping mechanisms.
- No consent to use scraped material.

3

On the other end of the sword are the issues that develop with your use AI to work on your project. Many of these issues impact the Doctrine of Fair use and your final material. READ

## ChatGPT as a Research Librarian/Researcher



According to CHATGPT AS RESEARCH SCIENTIST: PROBING GPT'S CAPABILITIES AS A RESEARCH LIBRARIAN, RESEARCH ETHICIST, DATA GENERATOR AND DATA PREDICTOR, <https://arxiv.org/pdf/2406.14765>

- GPT openly acknowledges that it is capability of providing fictional references, incomplete reference, e.g. lacking year, journal volume, or page number.
- On average, ChatGPT provides 62.41% hallucinated/fabricated information and 14.30% incomplete references
- In total, GPT-3.5 generated 39.05% more fictional references compared to GPT-4, which generated 23.12% .

4

Library Journal shared preprint article entitled CHATGPT AS RESEARCH SCIENTIST: PROBING GPT'S CAPABILITIES AS A RESEARCH LIBRARIAN, RESEARCH ETHICIST, DATA GENERATOR AND DATA PREDICTOR on June 20, 2024.

Though scientists widely adopt them, the promise of general-purpose artificial intelligence systems to facilitate science has been largely untested. In four studies, we examine the capabilities of ChatGPT across several tasks intrinsic to the scientific process. ChatGPT is a poor (but improving) curator of scientific articles. It is a surprisingly good research ethicist, detecting violations of statistical best practices and evolving open science protocols. Its ability to simulate known results may herald useful abilities in data generation and theory building. However, the chatbot had little success predicting highly novel data, highlighting its limited ability to surmise things outside its training data. Beyond merely testing LLMs, these studies produce several novel insights into the nature of machine intelligence.

In the study, they looked at how good a research scientist is ChatGPT? We systematically probed the capabilities of GPT3.5 and GPT-4 across four central components of the scientific process: as a Research Librarian, Research Ethicist, Data Generator, and Novel Data Predictor, using psychological science as a testing field. In Study 1 (Research Librarian), unlike human researchers, GPT-3.5 and GPT-4 hallucinated, authoritatively generating fictional references 36.0% and 5.4% of the time, respectively, although GPT-4 exhibited an evolving capacity to acknowledge its fictions. In Study 2 (Research Ethicist), GPT-4 (though not GPT-3.5) proved capable of detecting violations like p-hacking in fictional research protocols, correcting 88.6% of blatantly presented issues, and 72.6% of subtly presented issues. In Study 3 (Data Generator), both models consistently replicated patterns of cultural bias previously discovered in large language corpora, indicating that ChatGPT can simulate known results, an antecedent to usefulness for both data generation and skills like hypothesis generation. Contrastingly, in Study 4 (Novel Data Predictor), neither model was successful at predicting new results absent in their training data, and neither appeared to leverage substantially new information when predicting more versus less novel outcomes. Together, these results suggest that GPT is a flawed but rapidly improving librarian, a decent research ethicist already, capable of data generation in simple domains with known characteristics but poor at predicting novel patterns of empirical data to aid future experimentation.




## Cons of using AI in Research Reports

### Top 5 plagiarism statistics for 2025

- Plagiarism remains prevalent in academic settings, affecting **29%** of high school students and **28%** of college students.
- **59.7%** of content generated by GPT-3.5 included some form of plagiarism.
- **72%** of college professors familiar with ChatGPT express concerns about its role in cheating.
- **51%** of college students believe using AI tools like ChatGPT for schoolwork qualifies as cheating or plagiarism, yet around 1 in 5 still use them.
- The plagiarism checker software market is set to grow from **USD 90 billion in 2023 to USD 153.75 billion by 2031**.

5

On the other end of the sword are the issues that develop with you use AI to work on your project. Many of these issues impact the Doctrine of Fair use and your final material. READ



## Human v. Machine Authorship and Copyrightability

Human:

- Can be listed as an author
- Can copyright work
- Can submit peer review articles, documents, journals entries for publication
- No question of liability for plagiarism, copyright infringement, discrimination

Machine


- Cannot be listed as an author
- Must be identified and disclaimed in a document
- Cannot be copyrighted because it is not of a human creation
- Question of liability for plagiarism, copyright infringement, discrimination

6

When we talk about works generated by humans and machines we need to understand who can create what, who is an author, who is liable for infractions. READ

The Doctrine of Fair Use helps insure your creativity is safeguarded from others. The Doctrine now must balance between humans and machines.

Creativity is ideas and thoughts that come from our minds, and once we create something we hope that others will give us credit for that work. To create a work there must be a human factor. Generative AI does not make the cut. They are not as good as human in regard to input. AI may exhibit creativity, novelty, innovativeness but the question is whose creativity, novelty, innovativeness – the inventor of the AI tool and the data scrapped. Copyright law deals with the expression of the idea not the idea itself. AI cannot provide an expression of an idea, only the idea. Copyright also requires an author, artist, composer, etc. Authors can only humans, as such only a human can be an author, artist, composer, etc. Further, it is against the Doctrine of Fair use to copyright or author something that has already been copyrighted or written. AI has no way to gauge what came from someone else, what was previously copyrighted, or cannot tell, in most instances, where the information came from because AI tools are training on large datasets that can include copyrighted work without exclusive rights to make reproductions



## Safeguarding Data

- Disable chat histories and training when possible
- Refrain from uploading documents to LLMs
- Use plugins when necessary
- Only use AI programs on paraphrased sections or for grammar checking on one sentence at a time
- Do not check full sections or documents
- Prohibit ChatGPT and other programs for literature review purposes

7

To safeguard your creative works, you need to understand the AI program you are working with. You should look at safeguarding measures in the AI program.

Disabled “Chat History and Training”. By doing so, ChatGPT deletes all chat histories after 30 days and will not utilize user inputs for model training.

Opted out of having our data used to improve ChatGPT model performance. A form was submitted at [https://docs.google.com/forms/d/e/1FAIpQLScrnC-A7JFs4LbluzevQ\\_78hVERINqqCPct3d8XqnKOfdRdQ/viewform](https://docs.google.com/forms/d/e/1FAIpQLScrnC-A7JFs4LbluzevQ_78hVERINqqCPct3d8XqnKOfdRdQ/viewform)

Refrained from uploading documents to LLMs. Without specific plugins, ChatGPT cannot access PDFs anyway.

Restricted the use of LLMs to paraphrasing or grammar checking of only one sentence or two at a time, instead of providing an entire section/ paragraph.


Prohibited the use of ChatGPT for literature review purposes, given its tendency to generate fictional information and potentially imaginary references.

Please let me know if you’d like to discuss further.

Thanks,

Mohamed





## Opting Out of Data Scraping

- In ChatGPT: Go to settings, data controls, disable chat history and training.
- Google Gemini: You cannot completely disable or remove scraping tools. To limit scraping click on activity and select the turn off drop down menu. Data that is scraped will stay in the Gemini system for at least 3 years.
- Adobe: Deselect enable generative AI or share information feature, and click off allowing content to be shared under data and privacy settings.
- Grammarly: There is no opt out function.
- Install AI detection removal software or other bypass tools.
- Your company website: Install opt out features.

8

To safeguard your creative works, you need to understand the AI program you are working with. You should look at safeguarding measures in the AI program.

Disabled “Chat History and Training”. By doing so, ChatGPT deletes all chat histories after 30 days and will not utilize user inputs for model training.

Opted out of having our data used to improve ChatGPT model performance. A form was submitted at [https://docs.google.com/forms/d/e/1FAIpQLScrnC-A7JFs4LbluzevQ\\_78hVERINqqCPct3d8XqnKOfdRdQ/viewform](https://docs.google.com/forms/d/e/1FAIpQLScrnC-A7JFs4LbluzevQ_78hVERINqqCPct3d8XqnKOfdRdQ/viewform)

Refrained from uploading documents to LLMs. Without specific plugins, ChatGPT cannot access PDFs anyway.

Restricted the use of LLMs to paraphrasing or grammar checking of only one sentence or two at a time, instead of providing an entire section/ paragraph.

Prohibited the use of ChatGPT for literature review purposes, given its tendency to generate fictional information and potentially imaginary references.

Please let me know if you’d like to discuss further.

Thanks,

Mohamed

## Doctrine of Fair Use



The Copyright Act (Act) was codified in Title 17 of the United States Code, and in Chapter 1, Section 107 of the Act sets out the limitations on exclusive rights for copyrights. According to the Doctrine of Fair Use:

“...reproduction in copies or phonorecords or by any other means specified ..., for purposes such as criticism, comment, news reporting, teaching, scholarship, or research, is not an infringement of copyright.”

The Doctrine specifies that authors are allowed to use,

**“quotation of excerpts in a review ...; quotation of short passages in a scholarly or technical work, for illustration or clarification of the author’s observations; ... reproduction by a teacher or student of a small part of a work to illustrate a lesson; ...”**

9

The Doctrine of Fair Use allows for the free flow of creativity by “permitting the unlicensed use of copyright-protected works in certain circumstances.” (16) When writing DOT research reports, principal investigators, post doctorate and graduate students, as well as undergraduate students freely gather necessary information and produce peer reviewed journal articles written from research gathered from data that originates in DOT research projects. When drafting these publications, the authors must insure they do not infringe on any copyright, must insure they are not plagiarizing any content, and they must cite the information gathered from another cite. They also use AI programs to help write their final reports, whether it is Grammerly or a ChatBot, OpenAI, or other form of AI. What we must be wary of is the fact that AI models feel that because information is available to the public, then they should be free to use it to train their AI tools. AI program owners also feel that inputting/scanning of information, whether it comes from outside sources, or your research reports is not infringement. But the AI program may not keep track of where the information came from. AI companies claim that they do not copy original work but rather they digest it, in order to clear how human language functions work. They are learning from the data scraped and they are not responsible for the output of the data. Failure to cite/plagiarism, bias, inaccurate information, false references, etc is on the person using their services, not on the AI company even though they do not tell you where the information is coming from, whether it is copyrighted, and who owns the original information. The copyright alliance, authors, researchers and other believe that the scraping by AI programs does not allow for acknowledgement of the true author and can lead to plagiarism and infringement.



## Doctrine of Fair Use

The annotated section of the Copyright Act sets out uses and prohibitions of the use of copyrighted material as follows:

- "...the endless variety of situations and combinations of circumstances that can rise in particular cases precludes the formulation of exact rules in the statute. **The bill endorses the purpose and general scope of the judicial doctrine of fair use, but there is no disposition to freeze the doctrine in the statute, especially during a period of rapid technological change.** Beyond a very broad statutory explanation of what fair use is and some of the criteria applicable to it, the courts must be free to adapt the doctrine to particular situations on a case-by-case basis."

10

The federal government and the US Copyright office understand that there is a rapid technological change going on so they do not want laws to stop development, but they also do not want to see infringement issues arise. It's a balancing act that will be going on for a long time.

## Fair Use



The Doctrine of Fair Uses 4 prong test.

1. Purpose and character of the materials use-what work was used.
2. Nature of copyright work used-what source.
3. Amount and substantiality of work used-purpose.
4. The potential market for use.

11

The Doctrine was codified in 1976, in Section 107 of the Copyright Act in the manner of a 4 prong test. As we all know, authors will incorporate existing materials in their works with proper citations. Besides the 4 prong test, the Doctrine does not infringe upon fair payment-compensation for work; allows authors the right to distribute their work and where it is distributed; makes sure proper credit and attribution is given; and does not allow others to circumvent a persons ownership rights. This are points that some Generative AI companies are missing.

With the prongs a person/group, etc., are able to use works, including copyrighted works. AI companies will gather data and information claiming they are following the doctrine, and at no point is the original author given credit and copyrights are being infringed. This is what puts us at risk. When our contractors write reports and journal entries there is no way to know if the work they are citing is original work or if the information is work that is pulled from another's work that has not been cited. This can lead to plagiarism claims, infringement claims, and unfortunately this issues arise with no malice. It puts a stain on your company.

Prong one looks at commerciality and transformativeness of work. Transformative uses are "those that add something new, with a further purpose or different charter, and do not substitute for the original use of the work."

The second prong analyzes the degree to which the work that was used relates to copyright's purpose of encouraging creative express.(16) Technical articles compare and contrast existing information with conclusions being drawn, and appropriate credit must be given, licenses and permission requirement must be followed, and the author must be able to cite all sources.

Under the third prong, there is a need to look at the quantity and quality of the copyrighted material used. Individuals need to look at how much copyrighted work is being used in the training of and output from the AI tool, and the greater the portion, the more likely there is an infringement. A key component of whether there is fair use or not is whether the work generated from the AI tool takes the "heart of the original work's creative expression". (5) Further, the quality of work deriving from the output from an AI tool can be suspect due to unknown bias and inaccurate data that may exist in the AI tool's database. The issue for this prong arises because AI companies can now scrape data in real time and there is no way for the AI company to check for copyright, plagiarism, bias, or proper citations or references. In fact, references, citations and what have been deemed, "bogus judicial decisions...bogus quotes and bogus internal citations" (7) have cropped up due to the reliance on AI tools when drafting documents. In the medical, legal and many other communities, they are finding that the hallucination rate or rate of misleading information and bogus facts in peer-reviewed publications is on the rise.

Under the fourth prong, there is a need to look at the effect an AI product will have on the potential market for or value of the copyrighted material. Since there is a market for published work, and publishers and authors benefit from their copyrighted articles, there is a need to guard against AI tools scraping data to ensure that the final product that comes out of their data cannot and does not compete against the original work. This includes both the current work by an original author and any potential new works. Numerous court cases have reviewed whether a work from an AI-generated program that does not receive permission to use an original work and then benefits from the copied product falls outside of the Doctrine of Fair Use. The key determining factors are the potential market for the use of and the value of the copyrighted material.



## Publisher's Requirements

Authors who use AI tools in the writing of a manuscript, production of images or graphical elements of the paper, or in the collection and analysis of data, must be transparent in disclosing in the Materials and Methods (or similar section) of the paper how the AI tool was used and which tool was used. Authors are fully responsible for the content of their manuscript, even those parts produced by an AI tool, and are thus liable for any breach of publication ethics.

12

**From:** Randall, Claire E. <[CRandall@nas.edu](mailto:CRandall@nas.edu)>  
**Sent:** Tuesday, May 23, 2023 10:38 AM  
**To:** Randall, Claire E. <[CRandall@nas.edu](mailto:CRandall@nas.edu)>  
**Subject:** TRBAM Papers: Guidelines for use of Large Language Models (LLMs, like ChatGPT) and generative AI tools  
**Importance:** High

Please note that AI bots such as ChatGPT **should not be listed as an author** on your submission.  
Many thanks!

Patti  
Patti Lockhart  
Director of Publishing and Outreach  
Technical Activities Division, TRB  
The National Academies of Sciences, Engineering, and Medicine  
500 Fifth Street, NW | Washington, D.C. 20001  
202-334-2284 | [plockhart@nas.edu](mailto:plockhart@nas.edu) | [www.TRB.org](http://www.TRB.org)

Elsevier/science type digests  
Sage/TRR  
Multidisciplinary Digital Publishers  
The Council of Science Editors  
Cambridge University Press


Like publishers, researchers need to insure that the interaction with AI tools is set out in their final product so that stakeholder and those using the final product can judge for themselves whether the information can be trusted or not. All of our research revolves around being transparent and fair. According to the National Institute of Standards and Technology (NIST) AI systems are said to "be biased when it exhibits systematically inaccurate behavior." (9) To ensure that there is no bias in research reports, researchers need to be mindful of what data they are using, where the data comes from, and whether they can verify the data that is outputted by an AI tool.

Researchers need to keep in mind that the actual copying of information, whether by a human or a machine is seen by authors as one thing and by AI generated companies as something completely different. When reviewing the issue courts focus on knowledge and whether a human or a machine performs the copying. Unfortunately, when an AI tool gathers data, it does not mark which portions are copyrighted or who the original authors were. As such, students, professors, and researchers cannot verify where they obtained the information or if they are infringing on someone else's work. Because of this, determining liability in cases where plagiarism, unintentional bias, misleading information, and infringement is hard to prove. When using an AI tool, researchers should be sure to vet the data they gather with more scrutiny than they have had to in the past. Researchers cannot just hope the information they gather is verifiable.

Further, researchers need to be more diligent to ensure there are no substantial similarities between their original work and the work that was scrapped from by an AI tool. This has become harder with the advent of new and improved AI tools. Researchers need to focus on whether the outputted work they used can be traced to an original work. Similarities can be hard to prove, especially when using AI tools to write literature reviews and substantial pieces of research reports.

AI developers, designers, and companies replicate patterns in data and researchers now need to delineate between what work is written by a machine and what is written by a human. Having both human and machine written words in a document causes issues of accountability and liability. Nevertheless, the AI tool cannot be considered the author of the work, and as such, if there is no author and there is no accountability. For this reason, there must be a delineation between what the machine writes and what the human writes. In addition, there must be a determination in how much the human has control over the machine written information.

## Publisher's Requirements



Authors are required to:

- 1. **Clearly indicate the use of language models in the manuscript**, including which model was used and for what purpose. Please use the methods or acknowledgements section, as appropriate.
- 2. **Verify the accuracy, validity, and appropriateness of the content** and any citations generated by language models and correct any errors or inconsistencies.
- 3. **Provide a list of sources used to generate content** and citations, including those generated by language models. Double-check citations to ensure they are accurate and properly referenced.
- 4. **Be conscious of the potential for plagiarism** where the LLM may have reproduced substantial text from other sources. Check the original sources to be sure you are not plagiarizing someone else's work.
- 5. **Acknowledge the limitations of language models in the manuscript**, including the potential for bias, errors, and gaps in knowledge.

13

Elsevier/science type digests  
 Sage/TRR  
 Multidisciplinary Digital Publishers  
 The Council of Science Editors  
 Cambridge University Press

Like publishers, researchers need to insure that the interaction with AI tools is set out in their final product so that stakeholder and those using the final product can judge for themselves whether the information can be trusted or not. All of our research revolves around being transparent and fair. According to the National Institute of Standards and Technology (NIST) AI systems are said to “be biases when it exhibits systematically inaccurate behavior.” (9) To ensure that there is no bias in research reports, researchers need to be mindful of what data they are using, where the data comes from, and whether they can verify the data that is outputted by an AI tool.

Researchers need to keep in mind that the actual copying of information, whether by a human or a machine is seen by authors as one thing and by AI generated companies as something completely different. When reviewing the issue courts focus on knowledge and whether a human or a machine performs the copying. Unfortunately, when an AI tool gathers data, it does not mark which portions are copyrighted or who the original authors were. As such, students, professors, and researchers cannot verify where they obtained the information or if they are infringing on someone else's work. Because of this, determining liability in cases where plagiarism, unintentional bias, misleading information, and infringement is hard to prove. When using an AI tool, researchers should be sure to vet the data they gather with more scrutiny than they have had to in the past. Researchers cannot just hope the information they gather is verifiable.

Further, researchers need to be more diligent to ensure there are no substantial similarities between their original work and the work that was scrapped from by an AI tool. This has become harder with the advent of new and improved AI tools. Researchers need to focus on whether the outputted work they used can be traced to an original work. Similarities can be hard to prove, especially when using AI tools to write literature reviews and substantial pieces of research reports.

AI developers, designers, and companies replicate patterns in data and researchers now need to delineate between what work is written by a machine and what is written by a human. Having both human and machine written words in a document causes issues of accountability and liability. Nevertheless, the AI tool cannot be considered the author of the work, and as such, if there is no author and there is no accountability. For this reason, there must be a delineation between what the machine writes and what the human writes. In addition, there must be a determination in how much the human has control over the machine written information.



## Ethics with the use of AI

1. COPE, WAME, and JAMA, Elsevier, TRR and other publications and organizations ask for transparency and disclosures in a section of the paper which sets out how AI is used and which tools are used.
2. AI is not your voice but someone else's.
3. AI can harm your work and credibility.
4. AI can project unintended bias, factually incorrect information.
5. AI does not receive permission to use copyrighted material.
6. AI does not tell you where the original information comes from
7. The information obtained from AI may have errors, negative connotations, images may be inaccurate, mislabeled, and the information is probably distorted..

14

COPE means community of publication ethics

WAME means world association of medical editors