Before the
Department of Commerce
United States Patent and Trademark Office
Washington, D.C.

In the Matter of	)	
	)	Docket No. 180712626–8840–01
Request for Comments on	)	
Intellectual Property Protection for	)	83 FR 58201
Artificial Intelligence Innovation	)	

COMMENTS OF THE R STREET INSTITUTE

January 10, 2020

Prepared by:
Caleb Watney
Technology Policy Fellow
R Street Institute
1212 New York Ave NW, #900
Washington, D.C. 20001
(202) 525-5717
cwatney@rstreet.org

Introduction

On behalf of the R Street Institute (R Street), we respectfully submit these comments in response to the United States Patent and Trademark Office (USPTO) Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation.[1]

R Street is a free-market think tank that takes a pragmatic approach to public policy challenges.[2] R Street has written broadly about the importance of progress and competition in the development and application of artificial intelligence (AI)[3]. Given the rapidly advancing state of AI across many different domains, questions surrounding the intersection of intellectual property and algorithmic inputs/outputs are of vital importance.

Below is an area we would like to highlight as needing further study and consideration by the USPTO as AI applications continue to be deployed across the economy.

Clarify the fair-use exemption for AI training data

Imagine a hypothetical startup focused on creating a natural language processing application. One readily available source of human dialogue the company might consider using to train the application would be the last 50 years of Hollywood scripts, many of which are scrapable from various online databases. Such an endeavor, however, would stand on legally dubious grounds. These scripts remain copyrighted works, and there are no clear legal guidelines established to delineate what is allowable as fair use in machine learning (ML) training data and what is not.[4] More likely, this startup would avoid this potential legal minefield and consider what other, less risky datasets might be available.

This is the ambiguous state of copyright enforcement in ML today, and it is problematic. As legal scholar Amanda Levendowski has argued, the de facto privileging of frequently low-quality data that exist in the public domain (such as the Enron emails) has inadvertently biased the many AI applications that are built on them.[5]

---

[1] United States Patent and Trademark Office, *Intellectual Property Protection for Artificial Intelligence Innovation*, Request for Comments, Docket No. PTO-C-2019-0038, Oct. 30, 2019. https://www.federalregister.gov/documents/2019/10/30/2019-23638/request-for-comments-on-intellectual-property-protection-for-artificial-intelligence-innovation.

[2] See "About R Street," https://www.rstreet.org/about-r-street.

[3] See, e.g., Caleb Watney, "Reducing entry barriers in the development and application of AI," R Street Policy Study No. 153, Oct. 9, 2018. https://www.rstreet.org/2018/10/09/reducing-entry-barriers-in-the-development-and-application-of-ai; and "Comments of the R Street Institute to the Federal Trade Commission: The consumer welfare implications associated with the use of algorithmic decision tools, artificial intelligence and predictive analytics," Docket No. FTC-2018-0056, Aug. 15, 2018. https://www.rstreet.org/2018/08/15/comments-to-the-ftc-the-consumer-welfare-implications-associated-with-the-use-of-algorithmic-decision-tools-artificial-intelligence-and-predictive-analytics.

[4] While the traditional four factor copyright test is obviously still relevant, opinions vary as to how it might apply. See e.g., James Rosenfeld and Cydney Swofford Freeman, "Artificial Intelligence, Fair Use, and Using AI to Create New Works," *Davis Wright Tremaine*, Apr. 10, 2018. https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2018/04/artificial-intelligence-fair-use-and-using-ai-to-c and Benjamin Sobel, "Artificial Intelligence's Fair Use Crisis," *Columbia Journal of Law & the Arts*, Forthcoming, Sept. 7, 2017. https://ssrn.com/abstract=3032076.

[5] Amanda Levendowski, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem," *Washington Law Review* 93 (July 19, 2018), pp. 579-631. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3024938.

This reality may also have important and underexplored implications for the state of competition in AI. While large incumbent firms typically have available vast reams of consumer data that can be used to improve the performance of algorithmic tools, startups and smaller firms are more reliant on datasets scraped from the internet to help offset this advantage.[6] If startups can acquire a sufficient amount of relevant data, they can often launch a new product or service, which begets new data that they can use to maintain and improve their services. This virtuous cycle of sorts can help these firms compete with larger, more established ones.

An enormous number of copyrighted works are scrapable from the internet. These works could provide new arbitrage opportunities for scrappy startups willing to find and leverage interesting data sets. Indeed, considering the massive amount of data that might be included in these efforts, the full scope of what is possible admittedly difficult to fully grasp. Yet the data of these works are currently underexploited in part because of the legal ambiguities surrounding their use in ML.

Google has already showcased one use case for which this type of data might be leveraged. In 2016, a research division within Google used a corpus of 11,000 free e-books to show the potential improvements that could be made to a conversational AI program.[7] This effort sparked considerable controversy with groups like the Authors Guild who considered it a violation of the author's intended purpose and arguably a copyright violation.[8] Because this instance involved a research paper and was not used for commercial purposes, no suit was pursued. Notably, however, the original 'BookCorpus' dataset is no longer publicly hosted.[9]

If they choose, large incumbent firms like Google have the resources to fight these lengthy legal battles, given their significant legal teams. Startups and smaller companies, however, are far less likely to have these resources on staff. In practice, this means the current ambiguity surrounding the fair use exemption disproportionally hurts smaller firms.

Given the existing legal ambiguity and the significant potential benefits to be reaped, further study and clarification of the legal status of training data in copyright law should be a top priority when considering new ways to boost the prospects of competition and innovation in the AI space.

---

[6] Ryan Calo, "Artificial Intelligence Policy: A Primer and Roadmap," *UC Davis Law Review*, Aug. 9, 2017. pp. 424-25. https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Calo.pdf.
[7] Oriol Vinyals et al., "Generating Sentences from a Continuous Space," *Google Brain*, May 12, 2016. https://arxiv.org/pdf/1511.06349v4.pdf.
[8] See, e.g., Richard Lea, "Google swallows 11,000 novels to improve AI's conversation," *The Guardian*, Sept. 28, 2016. https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation.
[9] See, for example, a post that mentions the missing 'BookCorpus' dataset and gives instructions on how to recreate the dataset for oneself. Steven van de Graaf, "Replicating the Toronto BookCorpus dataset — a write-up," *Towards Data Science*, Dec. 6, 2019. https://towardsdatascience.com/replicating-the-toronto-bookcorpus-dataset-a-write-up-44ea7b87d091.

Conclusion

We appreciate the opportunity to comment on the USPTO's Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation and look forward to further participation in these discussions.

Respectfully submitted,

Caleb Watney
Technology Policy Fellow
R Street Institute