From: Scott Wilson <thequestionis@live.com>
Sent: Wednesday, November 13, 2019 5:18 PM
To: aipartnership
Subject: Re: Request for Comments on IP protection for AI innovation Docket No. PTO-C-2019-0038

It appears that companies such as Google and others are using copyrighted material for machine learning training datasets and creating for-profit products and patents that are partially dependent on copyrighted works. For instance, Google AI is using YouTube videos for machine learning products, and they have posted this information on their GoogleAI blog on multiple occasions. There are Google Books ML training datasets available in multiple locations on the internet, including at Amazon.com.

There are two types of machine learning processes that I am aware of - expressive and non-expressive. A non-expressive ML training dataset would catalogue how many times a keyword appears in a copyrighted text, and appears to be permitted by fair use. However, expressive ML training analyzing artistic characteristics of copyrighted works to create patented products and/or processes does not appear to be exempted by fair use copyright laws.

The whole issue revolves around the differences between research and product/profit creation in a corporate context. I am including an article below that examines these legal issues in more detail.

https://lawandarts.org/wp-content/uploads/sites/14/2017/12/41.1_Sobel-FINAL.pdf

or www.bensobel.org - first article on page links to the link above.

As a copyright holder, I am concerned that this is a widespread industry practice of using copyrighted material without the knowledge or consent of copyright holders to create for-profit AI products that not only exploits copyright creators, but increasingly separates them from the ability to profit from their own work. An example would be an AI music generator program that creates low cost music for soundtracks that was created using existing copyrighted works.

I have attempted to contact Google to determine if they are using my copyrighted content on YouTube for Machine Learning training datasets and so far there has been no response to my inquiry. I have not seen anything in their Terms of Service that claims secondary use rights to use copyrighted videos uploaded to

YouTube for Machine Learning training datasets, however they are definitely doing it. Even one of the founders of Google, Larry Page, has admitted in a TED Talk with Charlie Rose that their image recognition software that was trained on images of cats was derived from YouTube videos. This is not an open secret. It is a widespread industry practice that is taught in universities and at Google. To the best of my knowledge, the ethical issues associated with these industry standard practices of widespread data mining and website data scraping are not part of the curriculum.

Even archive.org is currently scanning copyrighted books despite lacking the permission of authors and publishers, and this will inevitably be used for ML training datasets without the knowledge or consent of copyright owners. There is no way to know if the millions of books scanned via Google Books/Project Ocean will eventually be used for the same purposes, as it would require very little effort to do so, and would be virtually impossible to prove or detect once it was initiated.

If you need additional confirmation of these claims please contact me for additional links. As an example I would suggest looking at the Download section of the website of ImageNet, which provides a list of image URLs that can be used for Machine Learning training datasets, but because some of the images may be copyrighted, they only provide the actual images at their own discretion for research purposes only. But by providing a dataset of URL links, they enable anyone to acquire potentially copyrighted content to be used for the creation of image recognition software and other for-profit products, without compensation, consent or knowledge of the original creators of the content. This is also a common practice for facial recognition software.

Another example is story generating software released this week by Open AI that was trained using the internet to generate an AI system that creates realistic fake news articles based partially on data seemingly acquired from news sites, which means copyrighted material. Since these tech companies are usually so secretive on the nature and content of these machine learning trading datasets, it is virtually impossible to determine what the sources of the ML training datasets are, which makes copyright infringement virtually impossible to document and enforce, since the algorithms and artificial neurons created as a result are essentially a black box that cannot even be audited by their creators in some cases. This is so widespread that it is leading to the bankruptcy of the creative industries, and must be remedied if the arts are going to survive in the coming years, as economic considerations will eventually override creative considerations, and could make creatives obsolete.

Ray Kurzweil, an employee of Google has recounted a conversation with Larry Page, the co-owner of Google, that the real purpose and intent of Google was not to create a search engine, but to create an Artificial General

Intelligence, owned and operated by Google, using all existing digitized knowledge to train it. Kevin Kelly of wired.com recalled a similar conversation with Larry Page, and both interviews are part of a documentary called "Google and the World Brain" (2013). Transcripts are available online of these first person accounts, and the documentary itself is available on Amazon Prime. In his book The Singularity is Near, Ray Kurzweil advocates for strong Intellectual Property (IP) protections because of these AI training data issues, however this is one of the few recommendations contained in his book that are not being implemented by Google in pursuit of their ambiguous AI goals. Jaron Lanier has also been a staunch advocate to change these uncompensated ML training data issues, and can speak with more clarity and background than any person in the tech industry in my opinion.

A similar investigation was announced yesterday by HHS regarding the widespread collection of electronic health records by Google AI / Google Cloud to create for-profit healthcare products such as Google's Cloud Healthcare API, which is a clear violation of existing HIPAA privacy laws, with penalties of up to 10 years in jail in some cases. Because of the huge economic clout of these companies, the ethical considerations of even examining medical records without patient knowledge or consent cannot be successfully litigated without government intervention into these obvious antitrust overreach activities by Google, Facebook, Amazon and Apple, among others.

If you look at the behavior of Google through the lens of their stated intent according to Larry Page, it is obvious as to what these various threads lead to. Virtually all of their free and paid services are oriented towards collecting all existing data, content and knowledge available worldwide, and Google.com, Gmail, YouTube, Fitbit, Google Cloud, Google Books, Google Play, Google Docs, Google Translate, Google Maps, G Suite, Cloud Healthcare API, LANDR, Android, reCaptcha, Google checking accounts, Google AI and DeepMind electronic health records sharing programs, etc. are all dedicated towards a singular goal - to be the first organization to reach Artificial General Intelligence.

It has been speculated upon by Ray Kurzweil, Nick Bostrom and others, that if Google is allowed to reach that goal without governmental oversight, even governments worldwide will be unable to stop them. Combined with Google's Quantum Supremacy research, they will inevitably reach their stated goal - to fundamentally transform the world.

To a large extent they have already succeeded at this goal, but if the YouTube Video Recommendation Engine is any indication, what the world is being transformed into does not appear to be a utopian vision at all.

Scott Wilson

youtube.com/metapunker