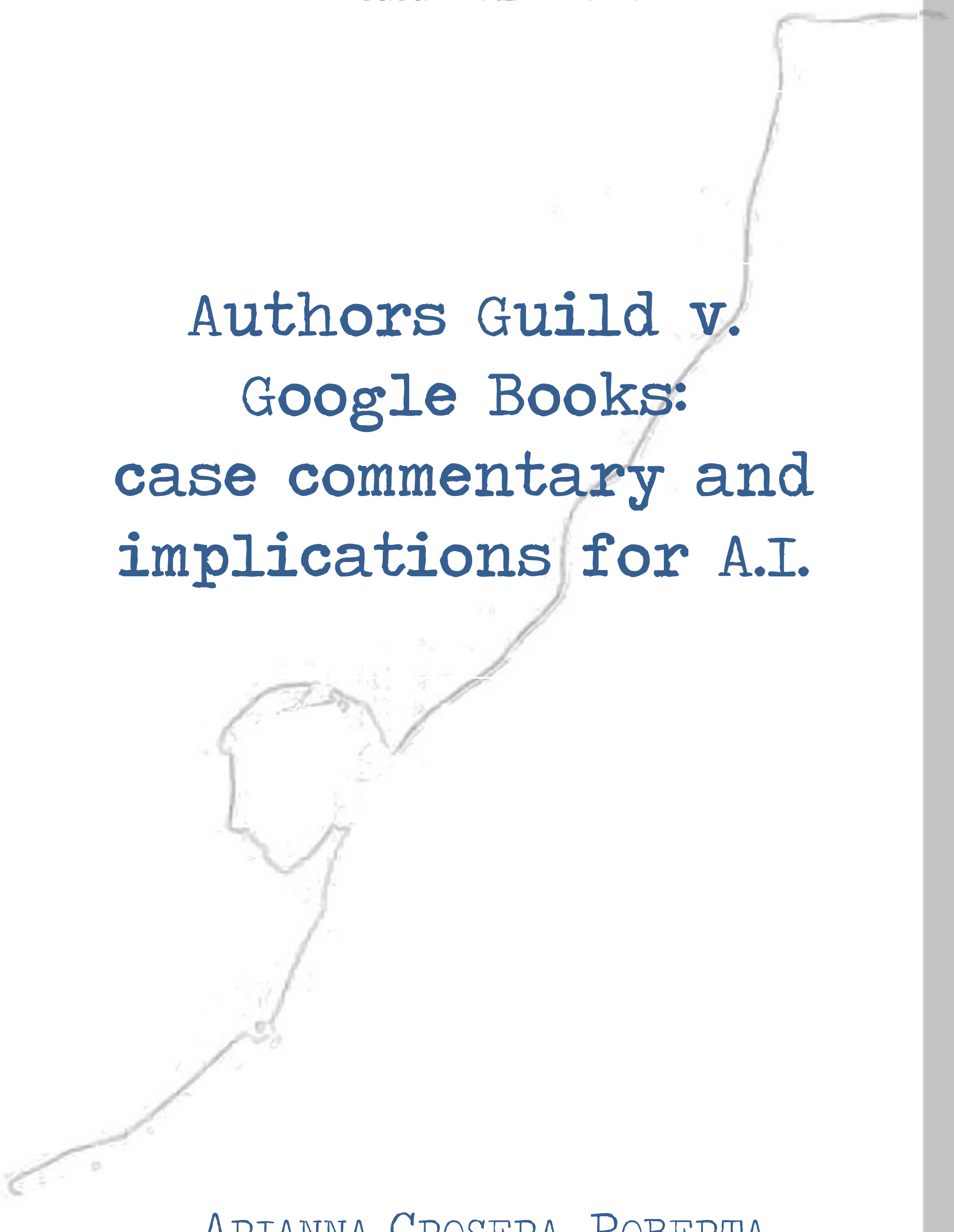
TRAIL Series

PAPER N. 2

a.a. 2019/2020

Authors Guild v. Google Books: case commentary and implications for A.I.



Arianna Crosera, Roberta Rombolà, Deborah Tosi, Darerca Tupponi, Alessia Zornetta

Trento BioLaw Selected Student Papers

Jean Monnet European Centre University of Trento

Co-funded by the Erasmus+ Programme of the European Union

Università di Trento

Facoltà di Giurisprudenza

I paper sono stati selezionati a conclusione del corso libero *Diritto e Intelligenza Artificiale* a.a. 2019-2020, organizzato all'interno del Progetto Jean Monnet “TrAIL – Trento Artificial Intelligence Laboratory”, coordinato presso l'Università di Trento dai docenti Carlo Casonato e Simone Penasa.

Authors Guild v. Google Books: case commentary and implications for A.I.

*Arianna Crosera, Roberta Rombolà, Deborah Tosi, Darerca Tupponi, Alessia Zornetta**

ABSTRACT: The subject of artificial intelligence is gradually affecting and spreading to all areas of law, including the one of intellectual property. This paper aims at analysing the role, importance and possible future of artificial intelligence in the field of copyright law, specifically through the analysis of the case Authors Guild v Google Books. The paper first offers a background of the facts, including an explanation of the technical aspects of the software, whose operation is the case of controversy. Secondly, it analyses the rationale behind the judicial proceedings and finally, it considers the deeper meaning behind the decision and the possible implications the judgement might have in the future.

KEYWORDS: artificial intelligence; intellectual property; generative algorithm; google books; case commentary.

1. Introduction

Artificial Intelligence is quickly evolving in society, and its fields of application are increasing significantly. The purpose of this paper is to demonstrate how the boundaries of artificial intelligence are stretching to all branches of the law, in this case Intellectual Property and especially the protection of copyrighted works. The case *Authors' Guild v. Google Inc*. is of particular interest because a simple process such as the scanning and classification of books, might not only infringe authors' rights but may also lead to the monopolization of the market through the unauthorized use of books to develop Artificial Intelligence mechanisms.

In section II. a background of the facts and a brief explanation of the algorithm is provided, followed by a summary of the judicial proceedings in section III. The main focus of the paper is found in section IV, where the implications of the case are discussed.

2. Background

As early as 1996, Google founders Sergey Brin and Larry Page had started to work on what is now Google Books[1]. In development since 2002, Google Print was launched in 2004 and was later renamed Google Book Search in 2005[2].

* The Authors are student of the Bachelor Comparative, European and International Legal Studies (CEILS), University of Trento.

[1] *Google Books History*, https://books.google.com/googlebooks/about/history.html (retrieved 05/02/2020).

The aim of its major initiative, the Google Books Library Project[3], was of digitalizing up to 15 million volumes within a decade and make them available through its service by partnering with universities and public libraries. Libraries such as those of the University of Michigan, Harvard University, Stanford University, Oxford University and the New York Public Library allowed Google to digitalize and render computer accessible innumerous books from their collection, a feat that would have been near herculean and extremely costly had they done so single-handedly, receiving as payment copies of the digitalized book for use in their own repositories.

Google not only digitalized books under public domain, but also those still under copyright, without seeking or obtaining permission from copyright holders[4]. This led to a major class-action lawsuit against Google in 2005 for copyright infringement by a group of authors and publishers led by the Authors Guild, followed by another lawsuit by the Association of American Publishers[5].

The Authors Guild's stand was not that of shutting down the Google Books service, but rather that of requiring Google to ask authorization to copyright owners and require that they be paid. At the time Google was monetizing its service, producing revenue for each search conducted on its Google Books browser, without having ever purchased a single copy of the books it was digitalizing, as well as utilizing this free data to train its algorithms which, in consequence, lead to a rising concern on Google obtaining a monopoly over search within digitalized books.[6]

The feature most alluring to the public using the Google Books service, especially those interested in academical research, was and still is the possibility to search for keywords within the text. Google Books provides a maximum of three snippets and a limited percentage of the content to browse in, previously approved by the copyright owner.[7]

In particular, Google Books' Ngram Viewer tool allows for searching and displaying how frequently a specific word or phrase can be found within books over selected years. This tool has proven to be a popular

[2] J. Grant., *Judging Book Search by its cover*, available at https://googleblog.blogspot.com/2005/11/judging-book-search-by-its-cover.html (retrieved 05/02/2020).

[3] *About the Library Project*, available at https://support.google.com/books/partner/answer/3398488?hl=en&ref_topic=3396243 (retrieved 05/02/2020).

[4] Authors Guild v. Google, Inc, 804 F.3d 202 (2d Cir. 2015).

[5] Authors Guild v. Google, available at https://www.authorsguild.org/where-we-stand/authors-guild-v-google/ (retrieved 05/02/2020).

[6] *Authors Guild v Google: Questions and Answers*, available at: https:// www.authorsguild.org/authors-guild-v-google-questions-answers/ (retrieved 05/02/2020).

[7] *An Introduction to the Google Books Partner Program*, available at https://support.google.com/books/partner/answer/3324395?hl=en&ref_topic=3238497 (retrieved 05/02/2020).

algorithm for researchers to analyze how concepts have emerged and developed through time[8], particularly to text mine in the field of language studies.

A single term can be analyzed, as well as multiple ones at the same time. Trends are shown in a graph through Ngrams addressing the percentage of use within a specific time span. Ngrams are constant sequences of characters like words, abbreviations and also numbers[9]. In its advanced usage results can be filtered through inflections, capitalization (case sensitive words) and part of speech (e.g., one word can be analyzed through its different uses as a noun, verb, adjective, etc.). It is possible to analyze trends within specific corpora, such as languages and geographical location[10].

Nonetheless, the Ngram Viewer has been the recipient of harsh criticism, wherein Google Books' database only represents a portion of existing publications, results between 1500 and 1800 are less reliable given the modest number of books published during that time frame[11], scientific literature seems to be overrepresented and all types of publications have the same weight and value[12].

3. Judicial Proceedings

Following the plaintiffs' claim of copyright infringement by Google, in 2008 a settlement was agreed on by the joint plaintiffs Authors Guild of America and Association of American Publishers, and Google Inc. According to the settlement agreement, Google was allowed to further scan and dispose of the books upon payment of $125 million dollars to cover the plaintiffs' court cost and to create a Book Rights Registry, and it had to keep on granting a widespread access to information.

However, on March 2011[13] the District court rejected the settlement because it would result in Google having a "significant competitive advantage" and being rewarded for the wrongful act of engaging in wholesale copying of copyrighted works without permission. It must also be noted that the conclusion reached by the court was heavily influenced by the significant number of critiques and reactions to the agreement, coming both from the United States and Europe[14].

[8] J. Chumtong, D. Kaldewey, *Beyond the Google Ngram Viewer: Bibliographic Databases and Journal Archives as Tools for the Quantitative Analysis of Scientific and Meta-Scientific Concepts*, in *FIW Working Paper*, n. 8, 2017, p. 7, available at https://www.fiw.uni-bonn.de/publikationen/FIWWorkingPaper/fiw-working-paper-no.-8 (retrieved 05/02/2020).

[9] J. Chumtong, D. Kaldewey, *op. cit.*, p. 6.

[10] *Ngram Viewer*, available at https://books.google.com/ngrams/info (retrieved 05/02/2020).

[11] M. Phillpott, *An Introduction to Text Mining*, available at https://port.sas.ac.uk/mod/book/view.php?id=554&chapterid=328 (retrieved 19/02/2020).

[12] S. Risi, *Google Ngrams: From Relative Frequencies to Absolute Counts*, available at http://stanford.edu/~risi/tutorials/absolute_ngram_counts.html (retrieved 20/02/2020).

[13] Authors Guild v. Google Inc., 770 F. Supp. 2d 666.

[14] A. Flood., *Ursula Le Guin leads revolt against Google digital book settlement*, available at https://www.theguardian.com/books/2010/jan/22/ursula-le-guin-revolt-google-digital (retrieved 18/02/2020).

An amended agreement was again rejected and on November 14, 2013[15] a second ruling by Judge Chin was issued dismissing the lawsuit. The argument in favour of Google was based on the fact that Google's use of the works was «fair», i.e. compliant with the copyright law.

When discussing Google's possible infringement of §106 of 17 U.S. Code[16], Judge Chin analyzed the doctrine of fair use. He firstly observed that the primary aim of copyright is to expand the knowledge and promote the Progress of Science and Useful Arts (quoting U.S. Const., Art. I, § 8, cl. 8). The main beneficiary is the public, whose access to knowledge copyright seeks to advance by providing rewards for authorship.

The fair use doctrine is codified in §107 of the Copyright Act. A case-by-case analysis needs to be done taking into consideration: (i) the purpose and character of the use (whether the use is of a commercial nature), (ii) the nature of the copyrighted work, (iii) the amount and substantiality of the portion used in relation to the work as a whole, and (iv) the effect of the use upon the potential market for or value of the work.

By making reference to the Campbell case of 1994[17], the court undertakes to explain the standards for finding fair use and differentiates between *transformative* and *non-transformative* use. Google's action is defined as transformative, since it digitalizes books and transform the text into a comprehensive word index with the aim of helping readers and others to find books.

According to Judge Chin, Google Books was not in violation of the copyright laws since, by facilitating access to the books, Google does not have a negative influence on the copyright holder. On the contrary, the search engine had the result of enhance the sales of the books, benefitting their authors.

The decision was appealed by the plaintiffs and reached the Second Circuit on April 11, 2014. Oral arguments were held on December 3, 2014, and continued until on October 16, 2015, the Second Circuit unanimously affirmed the judgment in Google's favour[18].

According to the court, Google's aim was «highly transformative» and did not result in a market substitute for the original works, yet, it increased the public knowledge by making information available about the authors works. Likewise, the fact that Google made the digital copies available to participating libraries is non-infringing since it required for libraries to use the copies in a way consistent with copyright law. On the third factor, the court confirmed the first instance court's finding that Google was not infringing the law because the text displayed was of limited amount. Regarding the fourth factor, the *snippet view* made it unlikely that the use of the search engine could significantly substitute the author's book as a whole.

[15] Authors Guild, Inc. v. Google Inc., 954 F. Supp. 2d 282m.

[16] 17 U.S.C. § 106 (2016).

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

[18] Authors Guild, Inc. v. Google Inc., No. 13-4829-cv (2d Cir. Oct. 16, 2015).

Notwithstanding the even stronger effects this ruling had on the precedent, the plaintiffs decided to file a writ of certiorari to the Supreme Court, but it was dismissed[19].

4. Implications

Authors Guild, Inc v Google, Inc is the expression that the way we see books has undeniably changed: the traditional perception of literary publications as mere works of art, aimed at spreading thoughts and information or simply narrating a fantasy has expanded into something more versatile.

Books have evolved into being tools used to train an algorithm in a given data set, or «raw ingredients to satisfy a machine's hunger for information»[20]. However, the problem with books is not merely a social one, but expands to the realm of law and, especially, intellectual property. The competent court explicitly declared that tech companies are legitimized to use copyrighted material such as, in this case, books, in order to train machine learning algorithms.

It's important to note that not all algorithms are the same. A relevant distinction is made between discriminative and generative models. Google Books is considered to be discriminative, which means that it takes the original data and essentially tries to break it down into a single result[21]. Instead, a generative algorithm creates data from the information it has collected. Since Google Books belongs to the first classification, it does not infringe fair use. Its use of data with the objective of creating a peculiar search engine and training its algorithm does not impact the financial earning of the authors of the books. However, as some experts sustain[22], there's a possibility that a similar reasoning might, in the future, be used as a precedent also for the generative model. This would create serious issues since such algorithm could, for instance, produce works similar to the ones of a specific artist of whom it has collected data. Such possibilities are already reality: a notorious example is The Next Rembrandt, a project whose development lasted eighteen months and whose final result was producing a painting extremely similar to the others by the Dutch artist from whom it takes its name[23]. The flowering of such programs leads to the following questions: who is to be considered the creator of the painting? Is it the research team behind the algorithm? What about the tech giants funding the project?

[19] Authors Guild v. Google, Inc., No. 15-849 (Dec. 31, 2015).

[20] R. Lea, *Google swallows 11,000 novels to improve AI's conversation*, available at https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation (retrieved 15/02/2020).

[21] M. Stewart, *The Most Important Court Decision For Data Science and Machine Learning*, available at https://towardsdatascience.com/the-most- important-supreme-court-decision-for-data-science-and-machine-learning-44cfc1c1bcaf (retrieved 18/02/2020).

[22] M. Stewart, *op.cit.*

[23] A. Winegar, *Protecting "The Next Rembrandt": Evaluating's Artificial Intelligence's Relationship With Copyright Law*, available at https://studentorgs.kentlaw.iit.edu/ckjip/protecting-next-rembrandt-evaluating-artificial-intelligences-relationship-copyright-law/ (retrieved 20/02/2020).

What if, instead, the algorithm can compose music? This is the aim of a program called "Bot Dylan", created by Dr. Obed Ben-Tal of Kingston University[24]. Dr. Ben-Tal is the head of the research team behind the Bot, so is he the one holding copyrights? Could the A.I. itself being considered the one with authorship? Therefore, questions arise, which are far too complex to discuss here: in the light of the obtainment of citizenships like Robot Sophia in Saudi Arabia, could Artificial Intelligence ever be considered the creative and original author of something? Secondly, could the mechanic and imposed work by algorithms be considered creative, being that an essential element for something to be considered copyrightable?

On a positive side, a precedent also opens a path for the legislators to enact regulations which can efficiently address issues which could arise from this situation, such as the distinction between what is discriminative and what is transformative in an algorithm[25]. Legislative action is also needed to avoid the risk of small companies, for instance startups, refraining from playing the same game as big businesses because they are afraid of breaching the law[26].

The case also opens another important issue: the role that powerful tech companies play in our information society. Their role is not confined to a mere technological and economic aspect: they are considered to possibly be beneficial for progress and preservation of culture. Indeed, Google's defence was that the algorithm not only is a search engine, but it also gives the possibility to bring back to life lost and out-of-print works and provides disadvantaged and rural-zones libraries with digital copies of books, thus ameliorating document searching.

Nevertheless, the advantages that such projects can offer have to be balanced with the risk of allowing single companies collecting a myriad of information. The danger is that companies such as Google might take the collected data and train their algorithm to make them more competitive than others, which would create a problem for competition law. And what could be the repercussions of a unilateral decision by Google to limit wholly or partially the content of their collection? It is sufficient to consider it as a mere hypothesis to understand that this gives tech companies an influential power on culture and knowledge. As the president of *Wikimedia Italia*, Andrea Zanni, explains: «Da una parte, l'istituzione americana rafforza il principio per cui tutti, studiosi e non solo, abbiamo il diritto alla conoscenza. E Google fa di più: non si limita a scansionare, fa anche text mining, cioè indicizza i testi e analizza i dati. Allo stesso tempo, però, delegare

[24] A. WINEGAR, *op.cit.*

[25] A clarification is needed here. The paper uses both the terms "transformative" and "generative" when referring to an algorithm. Such words have similar meaning and, when viewed from a general point of view, can be interchangeable. Both generative and transformative algorithm take an input and produce an output. The difference is drawn from the fact that a transformative algorithm usually produces an output of the same form of the input. Thus, a generative algorithm could also be transformative.

[26] A. WALKER, *Fair Use — The Exigency to Define the Law for AI Algorithms*, available at https://medium.com/datadriveninvestor/fair-use-and-ai-87f77721f1ea (retrieved 18/02/2020).

la digitalizzazione del sapere a una multinazionale rappresenta un rischio. E se Google a un certo punto "chiudesse" quel sapere? Se utilizzasse quei dati a proprio vantaggio competitivo? Ad esempio, Big G può mettere a frutto tutti quei testi per darli in pasto ai propri sistemi di intelligenza artificiale, in modo da collaudarli, "allenarli" ed essere più competitiva di altri. La questione va al di là del copyright: il 70% delle opere di biblioteca è "orfano", lasciare questo patrimonio in mano a una multinazionale significa anche darle una sorta di delega alla cultura"»[27].

Overall, the present case and its implications reflect the necessity to pair the development of technology with a regulatory framework able to define standards and limitations, especially for the conduct of the tech giants in the market.

This paper demonstrates that the recognition of the infinite possibilities that technology offers is not sufficient: legislators and judges need to be active in the creation and interpretation of the law in order to protect both the interests of civil society and the global market. Law-makers are called to apply their knowledge to issues related to the new technologies and artificial intelligence, engaging in a complex but absolutely necessary activity of revision of the existing norms and drafting of brand-new ones.

[27] V. Ertola, *La tutela del diritto d'autore ed il mondo digitale: Authors Guild of America vs Google*, available at https://www.iusinitinere.it/la-tutela-del-diritto-dautore-ed-il- mondo-digitale-authors-guild-of-america-vs-google-23271 (retrieved 19/02/2020).