2020

Copyright Law's Impact on Machine Intelligence in the United States and the European Union

Matthew Sag

COPYRIGHT LAW'S IMPACT ON MACHINE INTELLIGENCE IN THE UNITED STATES AND THE EUROPEAN UNION[‡]

Matthew Sag*

My topic today is copyright law's impact on machine intelligence and potential divergences between the law in the United States and the European Union. The defining feature of what we are calling the "Fourth Industrial Revolution" is an explosion of data. We need tools in order to navigate that data, and text data mining provides some of the most important to do so. Text data mining turns texts (or images or audio-video) into data. It is essential for any kind of search engine; it is essential for any kind of statistical analysis of large bodies of content; and it underpins a vast array of machine learning applications. Machines cannot "learn" unless the things they want to learn about have already been rendered into data, and that involves the practice of text data mining.

Machine learning is great for optimizing to a known objective. Machine learning is great for taking subjective human judgments and replicating them efficiently and at scale. So, one example that I like is that, people at MIT showed a computer system a bunch of photos and how the photos had been retouched by professional photographers.

The algorithm is now really good at retouching photos, right? It replicates subjective human judgment, yet in a new circumstance. It is not so great for producing nonfunctional objects. Maybe an AI can make a Rembrandt. It probably can take a photo and then give it a Rembrandt filter, effectively, but that software is not going to be radically creative. It is not going to jump paradigms and go to Greenwich Village and become a jazz musician.

What are the perils of AI? Well, we have a black box problem. The problem is that, although we can look under the hood, we will just not understand what we are seeing there. We also have a replication problem in that we cannot really necessarily replicate the algorithm if we do not have

[‡] This talk was given at the 2019 *FIU Law Review* Symposium: "Intelligent Entertainment: Algorithmic Generation and Regulation of Creative Works."

[*] Georgia Reithal Professor of Law, Associate Dean for Faculty Research and Development, Loyola University Chicago School of Law.

access to the data. And we also have an objectivity problem. What if the photo retouching software I described earlier makes everyone look whiter? What if it takes an existing social bias and replicates it? More peril. We face real anxiety that the current machine learning revolution will simply entrench and replicate the existing platform monopolies that we already have.

So, what does this have to do with copyright? These perils of AI and machine learning are not entirely copyright issues, but they have a significant copyright dimension. Copyright can actually exacerbate all of these problems because in a large number of scenarios, machine learning and AI rely on access to copyrighted works as inputs. If the reading machine is infringing copyright by the very act of reading, then we have a problem.

The problem is, twofold: we are not going to have enough data, and we are going to have an over-reliance on biased, low-friction data. For example, a lot of machine learning algorithms today use Wikipedia as a source. The people who write Wikipedia articles are not representative of the world population. Even in the U.S., they are not representative of the population. They skew nerd, they skew white, and they skew male. And so, whatever bias is in the training data is going to be replicated, codified, and sanctified with an undeserved aura of objectivity. If the reading machine is an infringing machine, then it is going to be very hard for new entrants to compete with the data that the large players have managed to aggregate already. It will also be difficult for researchers to replicate each other's findings, something that is a critical part of normal science.

The U.S. and the E.U. have taken quite different approaches to addressing the fundamental problems copyright seems to raise for text data mining, and thus by extension, machine learning. In the U.S., reproduction in the service of text data mining does not infringe copyright as long as the output of the TDM process itself is not itself infringing.

The U.S. approach is a triumph for fundamental principles and copyright, and it demonstrates the advantages of the flexible approach of the fair use doctrine. Fundamentally, American courts have understood and respected the most fundamental concept in copyright law, the idea-expression distinction. They have used that principle as the normative content of the fair use doctrine. The basic assumption in copyright is that copyright in the book does not give you copyright in the ideas or the facts embodied within that book. This has been the law since at least *Baker v. Seldon*.

The idea-expression distinction is an incontestable and fundamental part of U.S. copyright. And as far as I can tell, it seems to be an internationally accepted premise of copyright. We all know that original expression is required to make something copyrightable in the first place. But more than that, communicating original expression to the public turns out to be the hallmark of how we define all of the rights of the copyright owner.

In the U.S., courts have given life to this principle within the parameters of the fair use doctrine. They have looked at examples like software reverse engineering, plagiarism detection software, book search, meta-analysis of text, and they have seen that although these processes chew up a lot of copyrighted works as inputs, what they spit out the other end is either: (a) not at all copyrightable; or (b) if it is copyrightable, it is just a minor adjunct use like a snippet in a Google search result that is easily justified under traditional transformative use principles. Most importantly, in these examples, there is no expressive substitution for the original work, and that is why these uses have been held, over and over again, to be fair use.

In the E.U., they have taken a very different approach. This is probably because the whole European style of copyright law is different. The Europeans have the same understanding as we do about the distinction between ideas and their expression, but they have a radically different approach to thinking about copyright limitations and exceptions. That approach is much narrower, and dare I say, quite rigid. Prior to the Digital Single Market (or DSM) Directive, there was some scope to argue that some text mining could have been allowable under various E.U. exceptions. The quick summary is that those exceptions do not work very well, especially given the limited way that they have been interpreted.

Prior to the DSM Directive, a few European countries and the U.K. had enacted different types of text and data mining exceptions into their copyright law, although quite divergent systems of exception, which make it really hard to do anything cross-border in the E.U. With the DSM directive, we have the promise of a unified approach to text mining and copyright within the E.U. The DSM Directive is neither concise nor straightforward, and so my summary of it here is necessarily inexact.

Article 3 of the DSM Directive establishes its broad exception for text data mining by a somewhat narrowly defined class of nonprofit research organizations and cultural heritage institutions. Everyone who does not fall into that category has to look to Article 4. Your rights to conduct text mining under Article 4 are much more limited than under Article 3. They are limited chiefly in that rights holders can just opt out of text mining. Additionally, you can only do text mining in the worst possible way. You can mine the text and then instantly delete it as soon as you produce the metadata, which is not very practical for a lot of applications. If you want to replicate what you have done and say to people, "No, really, my results make sense. Here is an example," you cannot do that very easily once you have burned all the examples. That seems to be what E.U. law requires.

What I would like to talk about briefly is what we should make of these different approaches. I will begin by highlighting what some of the differences are.

In the U.S., fair use is for everyone. It is not reserved for special people or special classes of individuals. In the E.U., Article 4 is for everyone, and Article 3 is only for research organizations and cultural heritage institutions. Article 3 excludes commercial research, public-private partnerships, startups, and independent inventors. That is a massive problem.

In terms of the breadth of application, the fair use doctrine applies to all of the rights under the Copyright Act. It is very simple. In contrast, Article 3 and Article 4 primarily allow for reproductions and extractions. It is not at all clear that they apply to all the relevant rights under E.U. copyright law. E.U. copyright law has a right of adaptation, which is not mentioned in Article 3 or 4. E.U. copyright law has a right of communication to the public, which is not mentioned either. So, in the E.U., your ability to conduct TDM research and to use TDM as part of a machine learning process depends on exactly which rights you might be triggering in that process.

Finally, we come to the issue of contractual override. In the U.S., the fair use doctrine is not generally going to help you if you agreed in a contract not to do a specific thing. And so, if part of your license with one of the big publishers to get access to their material is a condition that you will not engage in text data mining, that is enforceable in the U.S. That is a huge problem, particularly if you want to do text mining in the medical literature, which is all wrapped up by the big publishing houses.

On the other hand, if you violate that license, their remedies sound in contract law, not copyright, and that matters a lot when it comes to the question of damages. In the E.U., the Article 3 exemption for nonprofit research organizations and cultural heritage institutions overrides any contractual restriction, which is pretty awesome from the point of view of researchers engaged in TDM and machine learning.

Similarly, in the U.S., your fair use right to engage in text data mining **does not give you a right to break down people's digital locks in order to do** your text data mining. In the E.U., if you are a nonprofit research organization and cultural heritage institution, it actually does, and that is pretty extraordinary. So, there are some instances that you would actually rather be a researcher in the E.U. than the U.S. if you face problems of contractual limitation or technological obstruction.

In the E.U., there is a proviso that you need to have lawful access to the underlying copyrighted works. That is just not an issue that we have faced yet in the U.S. It is a very interesting issue because of Sci-Hub. Retention of works is an issue I have already addressed. Under Article 3, the not-forprofits can retain those works. They can validate their results and they can do replication. But everyone else who relies on the Article 4 exemption cannot. I think that is a huge problem.

So, if we compare the U.S. to the E.U., we see a lot of similarities, but by and large, the United States' position is more flexible, particularly if you are not a nonprofit research organization and cultural heritage institution as that category is defined in the Directive.

Why do we have these differences?

The Europeans are just as smart as us, and I think their understanding of the essence of how copyright works is very similar to ours. But their understanding of why we have copyright is a little bit different. That is where some of the differences may come in.

The U.S. is much more utilitarian focused. The E.U. is much more natural rights focused, which leads them to a rigid system of exceptions and limitations that they think should be interpreted narrowly. And basically, a starting assumption that there is no exception. This translates into a different style or approach to the administration of copyright. In the U.S., we are fairly comfortable with letting judges do a lot of the initial heavy lifting through the fair use doctrine. Whereas in the E.U., it is much more a legislative approach.

These structural differences have now effectively opened up a substantive divide because in the U.S., I think we take the idea-expression distinction seriously. And in the E.U., the very fact that they have announced the need for these exceptions for text data mining indicates that they do not believe that it was otherwise lawful. And so, we could actually see a divergence on some of the basic substance of copyright there.

It is indisputable that the fair use doctrine gives the U.S. a technological advantage. Issues can be addressed as they arrive. Controversies can be resolved in accordance with core principles of copyright rather than by negotiating with and appeasing different interest groups. In the U.S., innovation proceeds litigation, whereas in the E.U., innovation lags regulation.

In the highly regulated system of E.U. copyright law innovation is always waiting for permission. And what confidence should an innovative firm have that permission will be forthcoming: in general, it is fair to say that the future has no lobby group. Of course, in 2019, Google has all the lobbying money in the world, but they had none back when they needed it the most (i.e., when they were starting the company).

It is not surprising that a lot of copyright and technology issues are addressed first in the U.S. through fair use. And then, other countries, basically, encode our case law judgments into their legislation. Now, the

Europeans might take issue with this characterization, but it is certainly true of some other jurisdictions.

Finally, I would like to turn to the commercial versus non-commercial distinction under the DSM Directive. This is essentially inexplicable in terms of copyright theory. The fact that it is so prominent in the E.U. shows us that there is a lot more than copyright that is driving the policy there. I do not want to suggest that it is a bad thing; it is just a different thing. Indeed, one could argue that the European approach has enabled them to actually take into account of non-copyright policy concerns. Some may characterize those concerns as protectionism, but I think there is a legitimate concern as we enter into the world of machine learning and AI that we are doomed to replicate existing power into these new markets. And so, one argument I could make in favor of the E.U. approach is that by giving a strong preference to the not-for-profit research sector, they might somehow avoid that trap. I do not know if I am convinced, but it is certainly an argument one could make.