Final paper: Nick Weninger

As many have noted and observed firsthand, one of the most significant ways that the constant presence of the Internet has shaped a majority of both societies and individuals is in its intricate and intimate connecting of them. This comes through in the rapid dissemination of information and conversation – the sharing of news articles, international interactions through social media, and other forms of online information sharing all allowing for different cultures and people to influence each other. At the same time, individual cultures, and more particularly seemingly isolated nations and sets of laws, can impact the way that the rest of the world receives and interacts with information through the Internet. This is particularly the case with the EU, which has passed legislation regarding the way websites and content providers are required to interact with consumers and content. Though these laws only apply to nations within the EU, countries outside of them are affected when international websites change their base policies and format to more easily comply with them. Most recently, the EU passed the highly controversial bill called the Copyright Directive (in full, the Directive on Copyright in the Digital Single Market), which has received warranted criticism for its potential to demand censorship and restrict free speech. In particular, the infamous "Article 13" will likely lead to widespread implementation of imperfect and problematic machine learning based filters that will censor Internet users and impact information accessibility worldwide (Fox 2019).

Ostensibly written with the goal of supporting creators in mind, the Copyright Directive, intended to be an overdue overhaul of EU copyright law, has been continuously met with backlash and protests from said content creators, EU citizens, and tech companies alike since it was introduced in 2016 (Reda 2019). Detractors often point to Article 11, which heavily restricts sites from posting excerpts, links, and screenshots of news stories, and especially Article 13, which has received the most media attention. Article 13, now listed as Article 17 after revisions (Browne 2019), makes service providers legally responsible for all copyrighted content uploaded to its platform and requires them to proactively prevent any illegal material from being uploaded or shared. Whereas previously copyright holders would have to reach out to service providers and request that copyrighted material be removed, with only those who uploaded the material being punished in the process, the Copyright Directive places the onus on the service providers, requiring them to proactive and not reactive to

potential copyright violations. As many have noted, this will inevitably lead to most service providers – including but far from limited to Facebook, YouTube, and other popular social media platforms – implementing broadly sweeping filters that will use black box machine learning algorithms to censor not only copyrighted content, but anything that is mistaken as such: including parodies, derivatives, wholly original content, and memes (Browne 2019; Doctorow 2019; Feiner 2019; Reda 2019). After extensive protests, the text of the policy was changed and the law's main supporter, Axel Voss, claims such filters are no longer required, but virtually everyone agrees that the change is meaningless and that companies are left with no alternatives if they wish to avoid violations and hefty fines. Regardless, as the laws have not yet been implemented in EU countries, neither have sites implemented the aforementioned, currently theoretical content filters or changed their site policies. However, assuming that the new laws are put into place as expected, past history and precedent can provide us with examples of both ways that users in other countries are impacted by Internet policies and laws of other nations, and how content filters might work, fail, and impact individual experience and freedom of expression.

This is certainly not the first time in recent years that a major EU policy regarding Internet regulations impacted citizens of the US and elsewhere. The General Data Protection Regulation (GDPR), a directive put in place May of 2018, is a set of policies intended to give Internet users more control over the way their personal data is managed. For example, Article 17 describes the "right to be forgotten," which requires data collecting organizations to provide users the option of requesting that their personal data be deleted - at least under certain conditions (Wolford 2018). It more heavily restricts the way that such data is initially collected and otherwise provides protections for the privacy of consumers. And even though the laws only apply nominally to residents of the European Union, users of many major international websites have been impacted as well (Shahani 2018). In 2018, many Americans – myself included – received emails and notifications from social media and shopping sites, informing them that their privacy policies had been updated. For many companies, it is simply more feasible to update their policies and the way they interact with users consistently around the world, instead of making different, modified versions of their platforms for users of different countries in addition to separating their collected data. Subsequently, it is important to remain abreast of international developments regarding Internet policy, as individuals may be more affected than they'd expect. In this case, the development is largely a positive one, and little to no machine learning is involved, but as we are seeing with the Copyright Directive, that is not always so.

Though most media coverage of the Copyright Directive does not go into any detail regarding the particulars of the machine learning involved in future content filters, past examples can again give u an idea of what to expect. In mid-2018, a set of bills under the name of FOSTA-SESTA were signed into law. Similar to the Copyright Directive, what was presented as long-needed reforming policies for a good cause – in this case, reducing the dangers of sexual exploitation – was beneath its shallow surface a set of rules that would do precisely the opposite and require websites to strictly filter their content. Like Article 13 of the Copyright Directive, the law places the onus on websites for the presence of any content violations; in this case, ads for any kind of sex work (Romano 2018). Due to the difficulty of accurately filtering such content, and unwilling to bear the risk, many websites simply removed entire sections of their site, as was the notable case of Craiglist's personals section, and others enacted broad new policies against nearly all content deemed sexual. In the latter case, image recognition machine-learning algorithms were implemented, in an attempt to flag and remove sexual content. Though likely created for additional reasons, an open-source pornographic content filter was developed and released by Yahoo, the owner of social-media platform Tumblr, in late 2016 (Mahadeokar and Pesavento, 2016). The model used is a variant of a convolutional neural network (CNN), a type of image identification machine learning where the input is pixels. Images are processed through multiple layers, with major features identified, colors assigned numbers, and non-linearity pulling out the differences that previous filters have noted and removing everything else (Živković 2018). Image recognition machine learning is typically supervised, with clearly defined labels and outcomes. In this particular model, images are processed and given a probability as an output, which indicates how likely the image is to be pornographic. In the 2016 study that was published, the primary objective seemed to be improving their model through help from outside sources. It seems unlikely that they were successful. In December of 2018 Tumblr banned and began removing content deemed pornographic through the use of image recognition machine learning – and whether it was the same CNN model used in the 2016 or another, it was clear that it was riddled with flaws and biases (Matsakis 2018). The filter tended to flag and remove excessive amounts of ostensibly allowed content, particularly art, LGBT+ posts, and posts discussing the filter itself. Many users experimented with their posts and showed how poor of a job the filter was doing, with many extreme examples including reposts of content published by Tumblr's own staff. The filter has certainly improved over time – however, this isn't necessarily a good thing. While it has improved, it's far from perfect, and ordinary content is still being removed, but at a rate small enough that people aren't protesting it anymore. All of this may be a portentous indication of the machine learning in our future when the Copyright Directive goes into

full effect in European countries, and content filters attempt to selectively and accurately remove images, text, and other content in far more broad categories.

It will be a few months, perhaps a couple years, before the EU's Copyright Directive is adopted into national law by the EU's member nations. In that time, it's possible additional policies could be passed, and that the adoption of the directive could add some nuance. But for Internet users in the United States and elsewhere, it is clear that the passage of major laws regarding content regulation has the potential to change and restrict the way we can communicate and share information online – particularly when machine learning is involved. Whether these changes happen now, in the future, or gradually over time, it is important to stay aware of these events and learn about the imperfect tools used to alter our daily lives. Image recognition models like convolutional neural networks may sound – and be – convoluted, but we can still work to improve our understanding of them and the ramifications they pose.

Citations

Browne, Ryan. 2019. What Europe's copyright overhaul means for Youtube, Facebook, and the way you use the internet. *CNBC LLC.* https://www.cnbc.com/2019/03/28/article-13-what-eu-copyright-directive-means-for-the-internet.html

Doctorow, Cory. 2019. The final version of the EU's copyright directive is the worst one yet. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2019/02/final-version-eus-copyright-directive-worst-one-yet

European Parliament. 2019. http://www.europarl.europa.eu/sides/getDoc.do?type=TA&language=EN&reference=P8-TA-2019-0231

Feiner, Lauren. 2019 YouTube and its users face an existential threat from the EU's new copyright directive. *CNBC.* https://www.cnbc.com/2019/05/10/youtube-faces-existential-threat-from-the-eus-new-copyright-directive.html

Fox, Chris. 2019. What is Article 13? The EU's copyright directive explained. *BBC News*. https://www.bbc.com/news/technology-47239600

Mahadeokar, Jay and Pesavento, Gerry. 2016. Open sourcing a deep learning solution for detecting NSFW images. *Yahoo Engineering*, *Tumblr*. https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for

Matsakis, Louise. 2018. Tumblr's porn-detecting AI has one job – and it's bad at it. *Wired.* https://www.wired.com/story/tumblr-porn-ai-adult-content/

Reda, Julia. 2019. The text of Article 13 and the EU Copyright Directive has just been finalised.

https://juliareda.eu/2019/02/eu-copyright-final-text/

Romano, Aja. 2018. A new law intended to curb sex trafficking threatens the future of the internet as we know it. *Vox*. https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom

Shahani, Aarti. 2018. 3 Things you should know about Europe's sweeping new data privacy law. *NPR*. https://www.npr.org/sections/alltechconsidered/2018/05/24/613983268/a-cheat-sheet-on-europe-s-sweeping-privacy-law

Wolford, Ben. 2018. Everything you need to know about the “Right to be forgotten”. *GDPR.EU*. https://gdpr.eu/right-to-be-forgotten/

Živković, Nikola. 2018. Introduction to convolutional neural networks. *Rubikscode.* https://rubikscode.net/2018/02/26/introduction-to-convolutional-neural-networks/