Osgoode Hall Law School of York University
Osgoode Digital Commons

All Papers

Research Papers, Working Papers, Conference Papers

4-14-2023

Do Automated Legal Threats Reduce Freedom of Expression Online? Results from a Natural Experiment

J. Nathan Matias

Merry Ember Mou

Jonathon W. Penney

Maximilian Klein

Lucas Wright

Follow this and additional works at: https://digitalcommons.osgoode.yorku.ca/all_papers

 Part of the Intellectual Property Law Commons

Do Automated Legal Threats Reduce Freedom of Expression Online? Results from a Natural Experiment

J. Nathan Matias[1*], Merry Ember Mou[1], Jonathon Penney[2], Maximilian Klein[1], and Lucas Wright[1]

[1]Cornell University Citizens and Technology Lab, Ithaca, USA
[2]Citizen Lab, University of Toronto; Osgoode Hall Law School, York U.
[*]nathan.matias@cornell.edu

ABSTRACT

Automated law enforcement systems support privately-operated enforcement bots to take legal action in hundreds of millions of cases a year. In the area of copyright, legal scholars have hypothesized the existence of "chilling effects" that harm public discourse by influencing people to self-censor protected speech. We test this hypothesis in a large-scale quasi-experiment with 9,818 accounts on Twitter that made 5,171,111 tweets. In a confirmatory interrupted time-series analysis, we find evidence that people reduce how much they post online after receiving a take-down notice from a copyright enforcement bot. On average, accounts sent fewer tweets after enforcement ($p<0.001$). Accounts also changed from a daily increase in public tweets to a decline on average ($p<0.001$). We also report on novel software that conducts third-party monitoring of the behavioral outcomes of automated law-enforcement systems. Since automated law enforcement can influence public discourse, third-party monitoring like this report will be essential to governing the power of enforcement algorithms in society.

Introduction

Automated systems routinely enforce laws hundreds of millions of times a year,[1] but the extent of their impacts are poorly understood.[2] These systems use continuous behavioral surveillance and computational models to automate legal actions against alleged offenders. Designers use automation to extend law enforcement into people's interpersonal lives with the goal of increasing the scale, responsiveness, efficiency, and consistency of law enforcement.[3] Yet even if these systems made perfect decisions, many scholars hypothesize that automated law enforcement could still deter law-abiding people from exercising fundamental human and constitutional rights.

When legal scholars and social scientists describe the deterrent role of law enforcement, they are understanding the legal system as a form of behavioral influence. In this view, a legal system is successful if it reduces unlawful behaviors but unsuccessful if it also prevents behaviors protected by basic rights. One mechanism for deterring behavior is through what psychologists call social norms—people's beliefs about what other people and institutions consider acceptable. When someone learns about, observes, or experiences enforcement, they incorporate that information into future decisions about how to act.[4] If people are less likely on average to engage in a behavior after becoming aware of law enforcement and surveillance, then the system has a deterrent effect.

Given the influence of legal systems on behavior, policymakers have been concerned about the side-effects of automated law enforcement on behaviors that are essential to a well-functioning democracy. A case in point is the U.S. Digital Millennium Copyright Right ("DMCA"), enacted in 1998 to respond to online copyright infringement. DMCA enforcement is now the most common form of automated law in the world. Under the DMCA, algorithms controlled by private enforcement firms routinely take hundreds of millions of legal actions each year.[5] Under the DMCA's "notice and takedown" regulatory scheme, copyright holders enforce their copyright by sending an automated copyright removal notice ("DMCA notice") to users and to online service providers (OSP) that host the infringing content. In parallel with debates over the accuracy of these private law-enforcement bots,[6] scholars have also expressed concern that DMCA enforcement might also deter forms of speech that are protected by law.[7] In this view, knowledge or experience of automated law enforcement causes people to fear enforcement and reduce their participation in public discourse overall.[8,9] If automated DMCA enforcement influences people against exercising their civil liberties, then it should be evaluated in light of that impact.

In this working paper, we present early evidence estimating the influence of automated law enforcement on the free exercise of civil liberties. In an interrupted time-series analysis, we observed the behavior of 9,818 Twitter accounts that received DMCA

takedown notices, before and after receiving the notice. In a cross-validated confirmatory study design, we then estimated the difference in the rate of tweets posted by those accounts before and after receiving the notice. While this quasi-experiment has some limitations, we expect this method and these findings to inform future causal work on the behavioral side-effects of automated law enforcement.

Automated Law Enforcement

Automated law enforcement refers to any system of legal action in which a computer is responsible for unsupervised decision-making. This may include the surveillance of behavior, analysis of data, and enforcement of the law.[10]. Automated law enforcement typically involves the use of software such as facial recognition, or software-enabled equipment as a camera or motion sensor.

When automation captures all three stages, from surveillance to enforcement, the role of human decision-making in the system is diminished, if not eliminated entirely. One widely adopted example of a fully automated cycle is traffic law enforcement via sensors and cameras that attempt to identify speed limit violations and traffic-light violations in order to automatically issue fines. Emerging examples of fully automated law enforcement include facial recognition cameras in China that detect jaywalkers, project their faces on billboards to shame them,[11] and can automatically issue a fine via SMS text.[12]

Automated law enforcement of speech is also possible wherever software systems mediate human communications.[13] While copyright enforcement was the earliest case of automated law enforcement online,[7] similar approaches are being applied in other areas of speech regulation. Recent laws passed in the European Union, such as Germany's Network Enforcement Act, require that social media platforms remove illegal content within 24 hours, incentivizing companies to widely-adopt fully automated moderation algorithms.[14]

In addition to wider concerns about surveillance and privacy, legal scholars have hypothesized four main harms from automated law enforcement.[10] First, automated systems could infringe on due process by violating the presumption of innocence and denying the right to face an accuser.[15] Second, software law enforcement can make systematic errors.[1] Third, automated decision-makers could reduce public perceptions of law enforcement fairness[16]. Finally, scholars worry that because automation enables high-volume, large-scale decision-making, existing problems with the legal system could be amplified and "reproduce inequality on a massive scale."[17]

In this paper, we consider a further risk from large-scale automated law enforcement: its potential deterrent effect on people's exercise civil liberties.

Automated Enforcement of The Digital Millennium Copyright Act

Under the DMCA's "notice and takedown" regulatory scheme, OSPs receive "safe harbor" protection from copyright liability in return for removing allegedly infringing content after receiving a copyright removal notice ("DMCA notice") from a claimed copyright holder. Once an OSP receives a DMCA notice and determines it is valid it must remove the alleged content and then also notify the user who posted the targeted content about the DMCA notice as well. Today, the number and scale of DMCA notices sent to OSPs in recent years has increased exponentially, largely due to automated law enforcement software. These algorithms scan the internet continually for copyright-infringing content and send on DMCA notices and removal requests to OSPs upon detection. In response, most major OSPs have built automated processes into their platforms to respond to these DMCA notices rather than have cases reviewed by human moderators.[18] In 2018, copyright enforcement firms sent over 700 million DMCA removal requests to Google.[5]

DMCA enforcement systems were pioneers of automated content moderation and policy enforcement and remain the leading forms of automated law enforcement online. Since the DMCA's enactment, platforms now conduct pro-active detection and removal of interpersonal threats, misinformation, and hateful language.[13] Other algorithms attempt to identify mental health crises in real-time.[19] Unlike the DMCA, these systems implement corporate policies unique to platforms that do not carry the weight of U.S. law. Posting racial hatreds and challenging their removal may result in temporary account suspension. Posting a clip from a famous cartoon and challenging its removal may result in a day in court and statutory damages up to tens of thousands of dollars.[1]

Chilling Effects From Automated Law Enforcement

Legal scholars have long hypothesized that copyright laws might have a "chilling effect" on freedom of expression. The idea of chilling effects, first used in a 1950s First Amendment Supreme Court Case,[2] has come to mean outcomes of policies that deter people from exercising their rights. As defined by Schauer, chilling effects occur where uncertainty in the legal system

[1] https://www.law.cornell.edu/uscode/text/17/504
[2] https://www.law.cornell.edu/supremecourt/text/344/183

creates risks and stimuli that deter people from behaviors that are protected as fundamental rights. For example, "a chilling effect occurs when individuals seeking to engage in activity protected by the first amendment are deterred from doing so by governmental regulations not specifically directed at that protected activity."[8] In recent years, scholars of internet law have hypothesized that both surveillance and copyright law exhibit these effects.[7,20]

In social psychological terms, chilling effects occur when people receive stimuli that update social norms, beliefs about what people and institutions consider acceptable.[4] If DMCA enforcement were perfectly accurate and people's understanding of that enforcement was also perfectly accurate, it would have no chilling effect on protected behavior unrelated to copyright. But if enforcement has room for error or people interpret the scope and risks too broadly, they might also self-censor by choosing not to engage in protected speech online.

Before automated law enforcement, most people did not need to worry about copyright regulations on speech and expression, nor would they likely ever face a copyright lawsuit. But automated law enforcement has the capacity to reach deeply into people's personal communications at an unprecedented scale, making hundreds of millions of highly error-prone legal decisions each year. In 2015, one research team estimated that 9.8% of all DMCA enforcement actions had some kind of error.[6]

Why might someone choose to self-censor in the future after receiving a take-down notice? First, copyright violators under the DMCA face civil penalties of $2,500 USD per violation or up to $25,000 USD in statutory damages. Criminally, violators face $500,000 USD in damages and up to 5 years in prison, with repeat offenders facing 1,000,000 USD in damages.[21] In light of these significant legal liabilities, a DMCA notice constitutes a significant and personalized legal threat, raising the likelihood of liability.[22] Second, receiving a DMCA notice alerts people about surveillance. Research has found that when people aware that they are being tracked, they are more likely to self-censor for fear of privacy or reputational harms.[22]

Since automated legal enforcement can be carried out at mass scale, harmful side effects can impact society as a whole, not just individuals. When people self-censor, they may not even notice the subtle shifts in their behavior to avoid scrutiny and shy from controversial but important topics.[9] Yet as these effects add up across society over time, they can result in weakened public discourse and a less informed democratic public. Furthermore, chilling effects can affect certain communities or groups disproportionately compared to others,[23] creating the potential for automated enforcement to "reproduce inequality on a massive scale".[17] Where these side effects occur, automated law's capacity to reach "into the public and the private spheres of citizens' lives" could "erode" trust between citizens and the state while "dehumanizing" the governing process.[10]

To test the hypothesis that automated content moderation by copyright enforcement software has a chilling effect on freedom of expression online, we created novel software to observe behavior on Twitter in real-time. We observed DMCA takedown notices as they were received by 9,818 people who use the social network platform Twitter, queried the history of their public activity for the 23 days before, and followed their activity on Twitter for the subsequent 23 days. Using this data, we are able to estimate whether people reduced public participation on Twitter after receiving a DMCA takedown notice.

Materials and Methods

To test the hypothesis of a possible chilling effect on expression from receiving copyright notices on Twitter, we collected data from the Lumen Database and matched it with observations of Twitter behavior. We then carried out a cross-validation process to develop and test a confirmatory hypothesis for the final analysis.

Data Collection

The Lumen Database is a research archive that records DMCA takedown requests submitted to Twitter, Google, Bing, and other online platforms. Using Lumen's research API, our software queried the Lumen database every 2 hours for the most recent posted DMCA take-down notices. Since individual take-down notices can mention more than one tweet or Twitter account, the software analyzed the text of the request for web addresses from *twitter.com* and extracted the unique Twitter accounts mentioned. We also recorded the date that the request was submitted to Twitter.

After identifying accounts that received take-down notices, our software determined whether an account was eligible for inclusion. Our software queried the Twitter API for more information about the account and labeled it for inclusion if the Twitter account reported their language as English (codes "EN" and "EN-US" were included, but "EN-GB" and others were excluded). If a Twitter account was mentioned in multiple DMCA notices, we recorded it and removed it from final analysis.

For accounts in the study, the software queried Twitter's historical API for as many recent tweets as possible, up to Twitter's limit of 3,200. Then, for the next 23 days, the software queried the Twitter API every 4 hours to observe new public tweets sent by those accounts. The software also took daily snapshots of account information, recording whether the account was permanently or temporarily suspended at the time. Accounts that were suspended for any reason were removed from the sample.

This study was reviewed and approved by MIT COUHES, the institution's research ethics review committee. Data was collected and processed in an encrypted datastore accessible only to the researchers.

	Full Data	Exploratory	Confirmatory
Twitter accounts	9,818	1,909	7,909
Total tweets	5,171,111	975,405	4,195,706
Earliest DMCA notice	2019-01-09	2019-01-09	2019-01-09
Latest DMCA notice	2019-02-13	2019-02-13	2019-02-13

Table 1. Descriptive statistics for the exploratory, confirmatory, and full datasets

Ultimately, the software collected a time-series dataset for every eligible participant that received a DMCA take-down request and whose account was accessible during the observation period. This participant-day dataset recorded the number of tweets an account sent in relation, as well as a day number indicating the number of days before or after receiving their DMCA notice (excluding the day of the notice). For each participant-day, whether it was a weekend or not and how many previous DMCA notices they had received during the observation period. In the final analysis, we excluded with langagues other than "EN" and "EN-US" and accounts that had received multiple DMCA notices. Because the Twitter API only guarantees the 3,200 most recent tweets, we also omitted accounts where the API returned more than 3,200 historical tweets, to avoid biases in the sample.

Cross Validation Procedure

The final sample included 5,171,111 tweets published between 2019-01-09 and 2019-02-13, from 9,818 Twitter accounts that received a DMCA takedown notice.

To carry out this study as confirmatory research, we split into two teams for cross-validation. The analysis team defined the research question and data collection procedures. The data controller team implemented the software and managed the data. After completing data collection, the data controllers provided a 20% exploratory dataset of 1,909 accounts and 975,405 tweets to the analysis team, who developed this paper and the accompanying analysis plan. Upon completion of this analysis plan, the data controllers shared the confirmatory dataset with the remaining 80% of accounts for analysis (Table 1). We report results from the exploratory, confirmatory, and full datasets.

Analysis

To study changes in the rate of public speech on Twitter after receiving a DMCA notice, we conducted a panel interrupted time-series study.[24] In this approach, we model the log-transformed number of tweets sent by an account in a single day, for each of the 9,818 accounts in the sample. Each of the 23 days before receiving the notice is recorded as a negative *DayNum* integer and each of the 23 days after receiving the notice is recorded as a positive integer. We omit observations on day 0, the day the report was filed. We record a binary variable *After* to indicate whether a given day was before or after receiving the notice. We also record whether the day was on a *Weekend* or not. Since different accounts publish tweets at different rates, we estimate the outcome using a random-intercepts maximum likelihood model,[24] with random intercepts for individual accounts.

$$log(y_{ij}+1) = \beta_0 + \beta_1 \times DayNum_{ij} + \beta_2 \times Weekend_{ij} + \beta_4 \times After_{ij} + \beta_5 \times After \times DayNum_{ij} + \varepsilon_{ij}$$

Results

On average, Twitter accounts that receive a single DMCA take-down notice reduce the number of tweets per day by 3.2 percent, a difference that is statistically significant ($p<0.001$). Before the notice, the trend is for accounts to increase the number of tweets they post per day. After receiving a DMCA notice, this trend reverses as people post fewer tweets over time ($p<0.001$, Figure 1).

Discussion

In this study, we find that on average, accounts reduce the number of tweets they post to Twitter by roughly 3% and also that they change from an increase in the number of tweets they post to a decrease on average, a reduction of 1.3% per day on average. When applied to the millions of people who receive these legal notices, the total impact on public discourse is likely substantial. As policymakers consider the future of automated law enforcement systems, they should also consider the impact of those systems on behaviors that exercise basic civil liberties.

Our findings are consistent with the hypothesis that DMCA enforcement deters future activity on Twitter, but the analysis in this working paper does not fully demonstrate the presence of a chilling effect on protected speech. Because this study only makes observations among people who received DMCA notices and does not include a comparison group, we cannot rule out

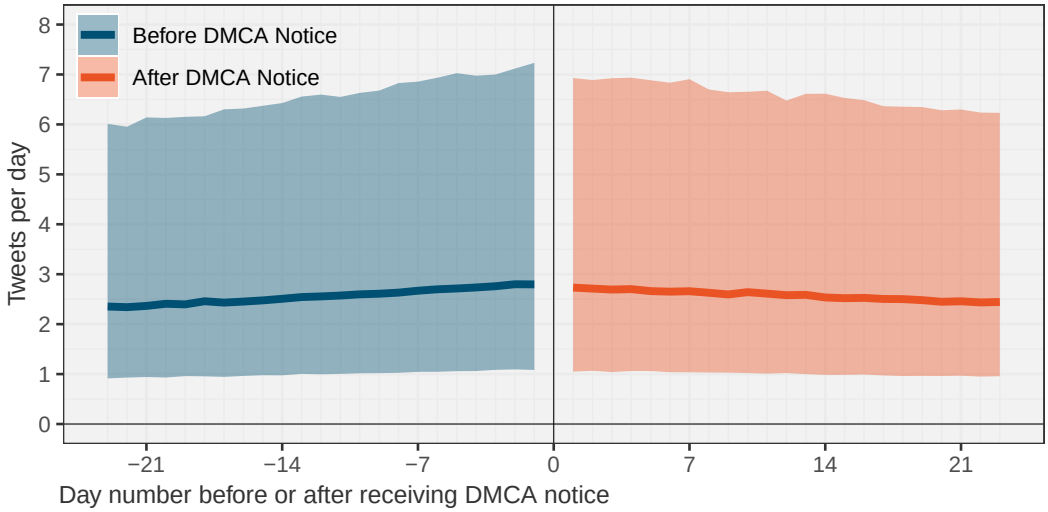

Figure 1. On average, accounts that received DMCA notices reduced the number of their tweets right away ($p<0.001$) and also reduced the rate of tweets over time ($p<0.001$) (n = 9,818 accounts & 5,171,111 tweets.)

alternative explanations for this change. For example, it is possible that the nature of the content people are posting (about a sporting event, for example) is such that their posting patterns would have declined anyway regardless of receiving the DMCA notice. Since we observe DMCA notices received over the course of five weeks and on many different days of the week, we do not expect that the differences we observe are due to wider seasonal trends over time. Further research could incorporate a comparison group to enable even more reliable causal estimates.

In this preliminary report, we have used language codes to limit our sample to accounts that we believe are largely based in the United States, this study likely included some accounts that are based elsewhere. A strict focus on the U.S. regulatory context would require more precise detection of US-based accounts.

References

1. Hartzog, W., Conti, G., Nelson, J. & Shay, L. A. Inefficiently Automated Law Enforcement. *Mich. State Law Rev.* **2015**, 1763–1796 (2015).
2. Hartzog, W. On Questioning Automation. *Cumberl. Law Rev.* **48**, 1–8 (2017).
3. Zittrain, J. Perfect enforcement on tomorrow's internet. In *Regulating technologies: Legal futures, regulatory frames and technological fixes*, 125–156 (Bloomsbury Publishing, 2007).
4. Tankard, M. E. & Paluck, E. L. Norm perception as a vehicle for social change. *Soc. Issues Policy Rev.* **10**, 181–211 (2016).
5. Van der Sar, E. Top 3 Copyright 'Owners' Sent Google a Billion Takedown Requests * TorrentFreak (2018).
6. Seng, D. K. B. 'Who Watches the Watchmen?'An Empirical Analysis of Errors in DMCA Takedown Notices. *An Empir. Analysis Errors DMCA Takedown Notices (January 23, 2015)* (2015).
7. Seltzer, W. Free speech unmoored in copyright's safe harbor: Chilling effects of the DMCA on the first amendment. *Harv. JL & Tech.* **24**, 171 (2010). Publisher: HeinOnline.
8. Schauer, F. Fear, Risk and the First Amendment: Unraveling the Chilling Effect. *Boston Univ. Law Rev.* **58**, 685–732 (1978).
9. Penney, J. W. Chilling effects: Online surveillance and Wikipedia use. *Berkeley Tech. LJ* **31**, 117 (2016). Publisher: HeinOnline.
10. Shay, L. A., Hartzog, W., Nelson, J., Larkin, D. & Conti, G. Confronting automated law enforcement. In *Robot Law* (2016). ISBN: 9781783476732 Publisher: Edward Elgar Publishing Section: Robot Law.
11. Mozur, P. Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras. *The New York Times* (2018).
12. Van Boom, D. Jaywalking in China? Surveillance system could SMS you a fine (2018).

	Full	Confirmatory	Exploratory
(Intercept)	1.335^{***}	1.334^{***}	1.335^{***}
	(0.012)	(0.028)	(0.014)
DayNum	0.008^{***}	0.008^{***}	0.008^{***}
	(0.000)	(0.000)	(0.000)
After	-0.032^{***}	-0.045^{***}	-0.029^{***}
	(0.004)	(0.009)	(0.005)
Weekend	-0.022^{***}	-0.017^{**}	-0.023^{***}
	(0.002)	(0.005)	(0.003)
DayNum x After	-0.013^{***}	-0.012^{***}	-0.014^{***}
	(0.000)	(0.001)	(0.000)
AIC	1085971.439	208191.474	877782.372
BIC	1086048.887	208257.448	877858.308
Log Likelihood	-542978.719	-104088.737	-438884.186
Num. obs.	471638	91581	380057
Num. groups: user.id	9818	1909	7909
Var: user.id (Intercept)	1.393	1.393	1.393
Var: Residual	0.529	0.513	0.533

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2. Results of analysis on the full, exploratory, and confirmatory datasets. The full analysis is the final result.

13. Gillespie, T. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press, 2018).

14. Bloch-Wehba, H. Automation in Moderation. SSRN Scholarly Paper ID 3521619, Social Science Research Network, Rochester, NY (2020).

15. Christensen, J. O. Wrong on Red: The Constitutional Case against Red-Light Cameras. *Wash. Univ. J. Law Policy* **32**, 443 (2010).

16. Wells, H. The Techno-Fix Versus The Fair Cop: Procedural (In)Justice and Automated Speed Limit Enforcement. *The Br. J. Criminol.* **48**, 798–817, DOI: 10.1093/bjc/azn058 (2008). Publisher: Oxford Academic.

17. Bornstein, S. Antidiscriminatory Algorithms. *Ala. Law Rev.* **70**, 519–572 (2018).

18. Gray, J. E. & Suzor, N. P. Playing with machines: Using machine learning to understand automated copyright enforcement at scale. *Big Data & Soc.* **7**, 2053951720919963 (2020). Publisher: SAGE Publications Sage UK: London, England.

19. Card, C. How Facebook AI Helps Suicide Prevention (2018). Section: Facebook.

20. Solove, D. J. A Taxonomy of Privacy. *Univ. Pennsylvania Law Rev.* **154**, 477–564 (2005).

21. The Digital Millennium Copyright Act of 1998. 18 (1998).

22. Penney, J. W. Privacy and Legal Automation: The DMCA as a Case Study. *Stanf. Technol. Law Rev.* **22**, 412 (2019).

23. Penney, J. W. Internet surveillance, regulation, and chilling effects online: a comparative case study. *Internet Policy Rev.* **6** (2017).

24. Singer, J. D. & Willett, J. B. *Applied longitudinal data analysis: Modeling change and event occurrence* (Oxford university press, 2003).

Acknowledgements

This research was supported by a grant from the AI Ethics Initiative.

Author contributions statement

M.M. developed the first proposed the study design and created a prototype system for collecting data. J.N.M. refined the study design, designed the research software, conducted the exploratory analysis, and served as lead author on the paper. J.P. informed the theory of the study, contributed to the analysis, and contributed to the writing. M.K. developed the production system used to collect this data and served as the data controller for the cross-validation process. L.W. contributed research and writing.