BLACK BOX TINKERING: BEYOND DISCLOSURE IN ALGORITHMIC ENFORCEMENT

Maayan Perel[*] *& Niva Elkin-Koren*[**]

Abstract

The pervasive growth of algorithmic enforcement magnifies current debates regarding the virtues of transparency. Using codes to conduct robust online enforcement not only amplifies the settled problem of magnitude, or "too-much-information," often associated with present-day disclosures, but it also imposes practical difficulties on relying on transparency as an adequate check for algorithmic enforcement. Algorithms are non-transparent by nature; their decision-making criteria are concealed behind a veil of code that we cannot easily read and comprehend. Additionally, these algorithms are dynamic in their ability to evolve according to different data patterns. This further makes them unpredictable. Moreover, algorithms that enforce online activity are mostly implemented by private, profit-maximizing entities, operating under minimal transparency obligations. As a result, generating proper accountability through traditional, passive observation of publicly available disclosures becomes impossible. Alternative means must therefore be ready to allow the public a meaningful and active interaction with the hidden algorithms that regulate its behavior.

This Essay explores the virtues of "black box" tinkering as means of generating accountability in algorithmic systems of online enforcement. Given the far-reaching implications of algorithmic enforcement of online content for public discourse and fundamental rights, this Essay advocates active public engagement in checking the practices of automatic enforcement systems. Using the test case of algorithmic online enforcement of copyright law, this Essay demonstrates the inadequacy of transparency in generating public oversight. This Essay further

* Dr. Maayan Perel, Post-doctoral Fellow, Haifa Center for Law & Technology, University of Haifa Faculty of Law; S.J.D., University of Pennsylvania Law School.

** Professor Niva Elkin-Koren, Director, Haifa Center for Law & Technology, University of Haifa Faculty of Law.

We thank Oren Bracha, Miriam Bitton, Jane Ginsberg, Ellen Goodman, Eldar Haber, Lital Helman, Joe Karaganis, Ethan Katsh, Shelly Kreiczer-Levy, Edward Lee, Neil Netanel, Gideon Pharchomovsky, Orna Rabinovitich-Einy, Tal Zarsky, as well as the participants of the 2015 Trust and Empirical Evidence in Law Making and Legal Process conference, University of Oxford, and the participants of the 2015 Conference on Empirical Research on Copyright Issues in Chicago-Kent College of Law, for their insightful comments and suggestions. Special thanks are due to the participants of the 2013–14 Law & Technology clinic, University of Haifa, for conducting the empirical study reported in this Essay, and to Nati Perel for the original design of the study. We are further grateful to Dalit Kan-Dror, Esq. for her academic assistance. This research was supported by the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation.

Electronic copy available at: https://ssrn.com/abstract=2741513

establishes the benefits of black box tinkering as a proactive methodology that encourages social activism. Finally, this Essay evaluates the possible legal implications of this methodology and proposes means to address them.

INTRODUCTION

We live in what Frank Pasquale named a "Black Box Society," wherein "[h]idden algorithms can make (or ruin) reputations, decide the destiny of entrepreneurs, or even devastate an entire economy."[1] The recognition that data-driven corporations play a growing role in determining opportunity and risk, basing their decisions on secret, "automated judgments that may be wrong, biased, or destructive," has

1. *The Black Box Society: About This Book*, HARV. U. PRESS, http://www.hup.harvard.edu/catalog.php?isbn=9780674368279 (last visited Oct. 3, 2016).

Electronic copy available at: https://ssrn.com/abstract=2741513

emerged in recent academic literature.[2] But digital devices and networked infrastructures,[3] powered by proprietary algorithms, control aspects of our everyday lives beyond money and information.[4] Increasingly, "black box" algorithms also rule legal regimes of law enforcement. Through various systems of automatic law enforcement, algorithms detect speeding (through red-light cameras),[5] prevent criminal activity (with the aid of GPS-enabled bracelets or anklets that alert whenever an offender enters a prohibited area),[6] or block allegedly infringing online content (relying on content filtering technologies).[7]

Algorithmic law enforcement is ubiquitous online, where behavior is inherently mediated by computer codes.[8] Indeed, algorithms substitute for the slow-to-react, occasionally burdensome legal system an efficient means to manage, organize, and analyze today's massive amounts of online data with uniformity and particularity, and to structure decision-making accordingly.[9] Still, algorithms are not perfect enforcers. Scholars have pointed out that algorithms may occasionally reach incorrect,

2. FRANK PASQUALE, THE BLACK BOX SOCIETY 18 (2015); *see also* ROBERT STOWE ENGLAND, BLACK BOX CASINO 2–3 (2011) (discussing how black boxes in the banking industry are on the rise); Andrew W. Lo, *Reading About the Financial Crisis: A Twenty-One-Book Review*, 50 J. ECON. LITERATURE 151, 156–57 (2012) (discussing how the securitization of subprime mortgages contributed to the lack of transparency in the banking industry prior to the financial crisis of 2008). *See generally* ANDREW ROSS SORKIN, TOO BIG TO FAIL (2010) (discussing the lack of information available in the aftermath of the financial crisis of 2008).

3. *See* DAVE EVANS, THE INTERNET OF THINGS 3 (2011), http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf (noting that network equipment manufacturer Cisco predicts that there would be 25 billion networked devices in the world by 2015 and 50 billion by 2020).

4. Rob Kitchin, *Thinking Critically About and Researching Algorithms* 7 (The Programmable City, Working Paper No. 5, 2014), http://ssrn.com/abstract=2515786.

5. Section 316.0083, Florida Statutes, known as the Mark Wandall Traffic Safety Program, authorizes local governments to use red light cameras to enforce violations of sections 316.074(1) and 316.075(1)(c), both of which prohibit the running of red lights. FLA. STAT. § 316.008(8)(a) (2016); Chris Matyszczyk, *Tickets Issued Due to Red-Light Cameras Are Illegal, Says Florida Court*, CNET (Oct. 21, 2014), http://www.cnet.com/news/tickets-issued-due-to-red-light-cameras-are-illegal-says-florida-court/.

6. The Omnilink Corporation, for instance, offers this service for victims of domestic violence. *Omnilink*, AM. CORRECTIONS SPECIALISTS, http://americancorrections.com/omnilink.aspx (last visited Oct. 3, 2016).

7. *See How Content ID Works*, YOUTUBE, https://support.google.com/youtube/answer/2797370 (last visited Oct. 3, 2016).

8. Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 568 (1998); *see also* LAWRENCE LESSIG, CODE 4 (2006) (discussing the emergence of cyberspace as a sphere of control).

9. *See* Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669, 687–88 (2010).

Electronic copy available at: https://ssrn.com/abstract=2741513

unjustified, or unfair outcomes,[10] especially when employed by private profit-maximizing actors.[11] Therefore, algorithmic decision-making cannot escape meaningful scrutiny. Proper accountability mechanisms are vital for policymakers, legislators, courts, and the general public to check algorithmic enforcement.[12] Yet algorithmic enforcement largely remains a black box. It is unknown what decisions are made, how they are made, and what specific data and principles shape them.

Normally, with human decision-making, oversight is principally achieved through transparency—so much so that the terms "transparency" and "accountability" are often used interchangeably.[13] In the realm of algorithmic enforcement, however, transparency alone is insufficient to generate accountability, for algorithms—due to their inherent traits—lack critical reflection.[14] First, algorithmic decision-

10. *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 673 (2016); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1256 (2008) ("The Colorado Benefits Management System (CBMS) has issued hundreds of thousands of incorrect Medicaid, food stamp, and welfare eligibility determinations and benefit calculations since its launch in September 2004. Many of these errors can be attributed to programmers' incorrect translations of hundreds of rules into computer code." (footnote omitted)).

11. Edward Lee, *Recognizing Rights in Real Time: The Role of Google in the EU Right to Be Forgotten*, 49 U.C. DAVIS L. REV. 1017, 1073 (2016); *see infra* Section II.B.

12. Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 531–32 (2016).

13. For instance, as part of President Barack Obama's stated quest to enhance the transparency of federal agencies, he signed the Transparency and Open Government Memorandum, declaring that he was "committed to creating an unprecedented level of openness in Government" and that he aimed to "promote[] accountability and provide[] information for citizens about what their Government is doing." Memorandum from the White House to the Heads of Exec. Dep'ts & Agencies (Jan. 21, 2009), 74 Fed. Reg. 4685, 4685 (Jan. 26, 2009); *see also* ADRIAN VERMEULE, MECHANISMS OF DEMOCRACY 182 (2007) ("Transparency is necessary for accountability, and helps to promote impartiality by suppressing self-interested official behavior."); Mark Fenster, *Seeing the State: Transparency as Metaphor*, 62 ADMIN. L. REV. 617, 619 (2010) ("To be held accountable and to perform well, [the government] must be visible to the public."); Mark Fenster, *The Opacity of Transparency*, 91 IOWA L. REV. 885, 900 (2006) [hereinafter Fenster, *Opacity of Transparency*] ("[Transparency] enables the free flow of information among public agencies and private individuals, allowing input, review, and criticism of government action, and thereby increases the quality of governance."); Adam M. Samaha, *Government Secrets, Constitutional Law, and Platforms for Judicial Intervention*, 53 UCLA L. REV. 909, 917 (2006) ("[P]opular accountability need[s] a system for disclosing information about government."); Frederick Schauer, *Transparency in Three Dimensions*, 2011 U. ILL. L. REV. 1339, 1346 ("Foremost among [the aims of transparency], at least in much of contemporary discourse, is what is commonly described as 'accountability.'").

14. Kitchin, *supra* note 4, at 7; *see also* Julie Brill, Former Comm'r, Fed. Trade Comm'n, Keynote Address Before Coalition for Networked Information 8–9 (Dec. 15, 2015), https://www.ftc.gov/system/files/documents/public_statements/895843/151216cnikeynote.pdf (where former FTC Commissioner, Julie Brill, acknowledged the challenge in creating public-facing

Electronic copy available at: https://ssrn.com/abstract=2741513

making is essentially concealed behind a veil of a code, which is often protected under trade secrecy law, and even when it is not, its mathematical complexity and learning capacities make it impenetrable.[15] Second, algorithmic enforcement is becoming so pervasive that transparency about the inputs and outputs of the algorithmic decision-making criteria may produce immense volumes of unintelligible data.[16] Without proper tools to analyze massive amounts of data, these overwhelming disclosures are mostly pointless. Third, when transparency is voluntary, as it often is with algorithmic enforcement implemented by private actors, the data disclosed may be partial, biased, or even misleading.[17]

Given the transparency shortcomings of algorithmic enforcement, black box tinkering becomes an important tool for generating social activism as a check on algorithmic governance.[18] Black box tinkering is a reverse engineering technique: a "process of articulating the specifications of a system through a rigorous examination drawing on domain knowledge, observation, and deduction to unearth a model of how that system works."[19] In the context of algorithmic enforcement, the ability to challenge the regulating code and confront it with different scenarios can reveal the blueprints of its decision-making process.[20] Put more simply, black box tinkering enables individuals to interact with the hidden algorithms that regulate their behavior.

Accordingly, this Essay explores the virtues of black box tinkering in promoting accountability in algorithmic law enforcement. It further demonstrates the benefits of this methodology in the context of algorithmic copyright enforcement, relying on the findings of a recent black box tinkering study striving to investigate online copyright enforcement practices by online intermediaries. It focuses on online algorithmic copyright enforcement for two main reasons. First, it has become ubiquitous, embedded in the system design of all major intermediaries as algorithms are used to monitor, filter, block, and disable access to allegedly infringing content.[21] Such a robust online infrastructure of algorithmic law enforcement, in the hands of a small

---

algorithmic transparency, calling on companies to proactively look internally to identify unfair, unethical, or discriminatory effects of their data use).

15. *See infra* Section I.A.

16. *See infra* Section I.C.

17. *See infra* Section I.B.

18. *See* NICHOLAS DIAKOPOULOS, COLUMBIA JOURNALISM SCH., ALGORITHMIC ACCOUNTABILITY REPORTING: ON THE INVESTIGATION OF BLACK BOXES 12–14 (2013), http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Algorithmic-Accountability-Reporting_final.pdf.

19. *Id.* at 13.

20. *Id.* at 14.

21. Perel & Elkin-Koren, *supra* note 12, at 480.

Electronic copy available at: https://ssrn.com/abstract=2741513

number of private and possibly biased megaplatforms, may create serious threats to the free flow of information.[22] Particularly, interested parties may abuse automatic systems of content adjudication to silence legitimate speech. [23] By engaging in black box tinkering, researchers can detect and subsequently contest undesirable censorship. Second, copyright enforcement involves a high degree of discretion, as many of the most serious issues in copyright law are extremely flexible.[24] A black box tinkering methodology is especially crucial for extracting valuable information about the way discretional standards are effectively translated into computerized codes.[25]

The Essay proceeds in three Parts. Part I explains why transparency is insufficient to generate accountability in algorithmic enforcement, using online copyright enforcement as a case study. Part II demonstrates the benefits of black box tinkering, using the example of a recent study of online copyright enforcement practices by online intermediaries. Part III raises possible legal challenges that may discourage researchers and activists from applying black box tinkering methodologies and suggests how to address them.

I. THE NEED TO STRIVE BEYOND TRANSPARENCY IN ALGORITHMIC ENFORCEMENT

> "We should interrogate the architecture of cyberspace as we interrogate the code of Congress."[26]

Traditionally, transparency has been conceived as the principal safeguard for human-driven regulatory accountability. It is generally assumed that public knowledge of the details of exercising governmental powers can counter abuse of power and dysfunctional governance.[27] President James Madison, for instance, famously stressed that a "popular [g]overnment, without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy; or, perhaps both."[28] A century later, Justice Louis Brandeis noted that "[s]unlight is said to be the best

22. *Id.* at 21; John Tehranian, *The New ©ensorship*, 101 IOWA L. REV. 245, 252 (2015). For examples of censorship caused as a result of third parties abusing the algorithmic system of N&TD, see Perel & Elkin-Koren, *supra* note 12, at 489–90.

23. Perel & Elkin-Koren, *supra* note 12, at 491–92.

24. *See infra* note 91 and accompanying text.

25. *See infra* Section I.D.

26. Lawrence Lessig, *Code Is Law*, HARV. MAG. (Jan. 1, 2000), http://harvardmagazine.com/2000/01/code-is-law-html.

27. *See* sources cited *supra* note 13.

28. Letter from James Madison to W. T. Barry (Aug. 4, 1822), *in* 9 THE WRITINGS OF JAMES MADISON 103 (Gaillard Hunt ed., 1910).

Electronic copy available at: https://ssrn.com/abstract=2741513

of disinfectants; electric light the most efficient policeman."[29] Even at the dawn of the twenty-first century President Barack Obama stated that a "democracy requires accountability, and accountability requires transparency."[30] The FTC has stated: "It is a basic tenet of our economic system that information in the hands of consumers facilitates rational purchase decisions; and, moreover, is an absolute necessity for efficient functioning of the economy."[31] Even when algorithms began replacing human judgment in performing administrative duties, some scholars continued advocating for enhanced transparency as a means of facilitating oversight.[32]

Nevertheless, this common perception of transparency as the ultimate guardian of decision makers in modern democracies is increasingly being challenged.[33] As Professor Mark Fenster argues, regulatory transparency is costly, impedes law enforcement and security objectives, and inhibits "the ability of government officials to deliberate over policy matters . . . without the inevitable pressure that accompanies public scrutiny."[34] Besides, it is unclear how regulatory transparency would be practically achieved—"what types of regulatory information should be made public, how this information should be presented, and how the potential pitfalls of transparency should be avoided."[35] Consequently, it is questionable whether transparency alone can elicit the type of public outcry that would compel an agency to change its course of action.[36]

This Essay argues that the current shift toward algorithmic governance, especially when performed on private grounds, bolsters the pitfalls of counting on transparency to generate proper oversight.[37] In the

29. LOUIS D. BRANDEIS, OTHER PEOPLE'S MONEY AND HOW THE BANKERS USE IT 92 (1914).

30. Memorandum from the White House to the Heads of Exec. Dep'ts & Agencies (Jan. 21, 2009), 74 Fed. Reg. 4683, 4683 (Jan. 26, 2009).

31. Proprietary Vocational and Home Study Schools, 43 Fed. Reg. 60,796, 60,805 (Dec. 28, 1978) (to be codified at 16 C.F.R. pt. 438).

32. *See* PASQUALE, *supra* note 2, at 10–12; DENA CHEN ET AL., PUB. KNOWLEDGE, UPDATING 17 U.S.C. § 512'S NOTICE AND TAKEDOWN PROCEDURE FOR INNOVATORS, CREATORS, AND CONSUMERS 1–4 (2011), https://www.publicknowledge.org/files/docs/cranoticetakedown.pdf; Jennifer M. Urban et al., The Am. Assembly, *Notice and Takedown in Everyday Practice*, 49–52 (U.C. Berkeley, Pub. Law & Legal Theory Research Paper Series, Paper No. 2755628, 2016), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628.

33. *See generally* BRENNAN CTR. FOR JUSTICE, THE GOVERNING CRISIS: EXPLORING SOLUTIONS 1, 10, 26, 29 (2014), http://www.brennancenter.org/sites/default/files/publications/dysfunction%20V7%2005%2012.pdf; David Frum, *The Transparency Trap*, ATLANTIC (Sept. 2014), http://www.theatlantic.com/magazine/archive/2014/09/the-transparency-trap/375074 (claiming that transparency and accountability reforms in government "have weakened political authority . . . [and] yielded more lobbying, more expense, more delay, and more indecision").

34. Fenster, *Opacity of Transparency*, *supra* note 13, at 906–08.

35. Jennifer Shkabatur, *Transparency With(out) Accountability: Open Government in the United States*, 31 YALE L. & POL'Y REV. 79, 84 (2012).

36. *Id.*

37. *See* Brill, *supra* note 14, at 8.

Electronic copy available at: https://ssrn.com/abstract=2741513

following discussion, this Essay presents four reasons why transparency is inadequate to safeguard algorithmic accountability: (1) it is very difficult to read, follow, and predict the complex computer code that underlies algorithms; (2) transparency requirements are irrelevant to many private implementations of algorithmic governance that are subject to trade secrecy; (3) algorithmic governance is so robust that even without mandatory transparency it is impossible to review all the information already disclosed; (4) when algorithms are called on to replace humans in making determinations that involve discretion, transparency about the algorithms' inputs (the facts) and outputs (the outcomes) is not enough to allow adequate oversight. This is because a given legal outcome does not necessarily yield sufficient information about the reasoning behind it. The following Sections explain these justifications for going beyond transparency in algorithmic governance, while demonstrating their merits in the context of algorithmic copyright enforcement by online intermediaries.

A. *Algorithmic Enforcement Relies on Complex Code and Machine Learning*

The first reason that transparency alone cannot produce sufficient checks on algorithmic enforcement relates to two intertwined technical characteristics of algorithms. The first is their non-transparent nature, which makes it difficult to review their decision-making process.[38] While algorithms can be built to advance specific values and policies,[39] they are ultimately reduced to complex code that we (and most program

38. Citron, *supra* note 10, at 1254; Charles Vincent & Jean Camp, *Looking to the Internet for Models of Governance*, 6 ETHICS & INFO. TECH. 161, 161 (2004) (explaining that automated processes remove transparency); Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 PA. ST. L. REV. 285, 293 (2011). This is not to say, however, that algorithmic copyright enforcement merits a higher level of transparency than what manual copyright enforcement demands. Transparency "should be applied to all steps which might compromise rights of individuals and seem arbitrary, be they automated or manual. The level of automation needs not, on its own, merit a higher level of transparency." Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1552. Both regimes—the automated one and the human-implemented one—should eventually reach a similar degree of transparency, yet automated regimes, by their nature, inherently challenge this goal.

39. *See* Bruno Latour, *Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts*, *in* SHAPING TECHNOLOGY/BUILDING SOCIETY: STUDIES IN SOCIOTECHNICAL CHANGE 225, 225 (Wiebe E. Bijker & John Law eds., 1992); Helen Nissenbaum, *From Preemption to Circumvention: If Technology Regulates, Why Do We Need Regulation (and Vice Versa)?*, 26 BERKELEY TECH. L.J. 1367, 1373 (2011). For more discussion specifically with regard to copyright enforcement, see R. Polk Wagner, *Reconsidering the DMCA*, 42 HOUS. L. REV. 1107, 1108–10 (2005) (suggesting that the total regulatory effect combines both law and technology and that changes in technology affect the law and vice versa).

Electronic copy available at: https://ssrn.com/abstract=2741513

developers) cannot easily comprehend.[40] Deconstructing the final code requires knowing what cognitive frames—as well as social, political, economic, and legal motivations—shaped the programmers' choice.[41] Indeed, translating legal mandates into code inevitably embodies particular choices as to how the law is interpreted, which may be affected by a variety of extrajudicial considerations, including the conscious and unconscious professional assumptions of program developers, as well as various private business incentives.[42] Some disparity between the computational representation and the law as it operates in practice is therefore unavoidable.[43]

The second technical characteristic of algorithms, which makes transparency an inadequate tool for checking their practices, is their learning capacities. Machine learning is capable of identifying trends, relationships, and hidden patterns in disparate groups of data.[44] Many algorithms can adapt their code and shape performance based on experience.[45] For instance, Google and Facebook may run "dozens of different versions of an algorithm to assess their relative merits, with no guarantee that the version a user interacts with at one moment in time is the same as five seconds" earlier.[46] These learning capacities make algorithms much smarter than some other computer codes that operate by means of on–off rules.[47] So learning algorithms are not merely tools for implementing the goals of those employing them: they effectively shape the meaning of the goals themselves. As a result, transparent information about the structure of the underlying code may sometimes be relevant only to the precise moment when the information was originally released.

A brilliant metaphor, suggested by Professor Suresh Venkatasubramanian, demonstrates the unique essence of machine learning by analogy to recipes.[48] He compares a standard algorithm to a recipe that "takes 'inputs' (the ingredients), performs a set of simple

40. *See* Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L. 235, 246 (2011).

41. *See* Jay P. Kesan & Rajiv C. Shah, *Deconstructing Code*, 6 YALE J.L. & TECH. 277, 283 (2003–2004) ("STS [Science & Technology Studies] examines how technology is shaped by societal factors such as politics, institutions, economics, and social structures.").

42. Bamberger, *supra* note 9, at 675–76.

43. *See* Citron, *supra* note 10, at 1261–62; Harry Surden et al., *Representational Complexity in Law*, 11 INT'L CONF. ON ARTIFICIAL INTELLIGENCE & L. 193 (2007).

44. Bernhard Anrig et al., *The Role of Algorithms in Profiling*, *in* PROFILING THE EUROPEAN CITIZEN 65, 65 (Mireille Hildebrandt & Serge Gutwirth eds., 2008).

45. *See* Kitchin, *supra* note 4, at 8–9.

46. *Id.* at 16.

47. *See* Bamberger, *supra* note 9, at 676.

48. Suresh Venkatasubramanian, *When an Algorithm Isn't...*, MEDIUM (Oct. 1, 2015), https://medium.com/@geomblog/when-an-algorithm-isn-t-2b9fe01b9bb5.

Electronic copy available at: https://ssrn.com/abstract=2741513

and . . . well-defined steps, and then terminates after producing an 'output' (the meal)."[49] A *learning* algorithm, by contrast, is described as "a *procedure* for constructing a recipe."[50] Accordingly, a learning algorithm is "a game of roulette on a 50 dimensional wheel that lands on a particular spot (a recipe) based completely on how it was trained, what examples it saw, and how long it took to search."[51] The inputs and the outputs of a simple "recipe" algorithm are both quite easy to follow, which is not the case with smart, learning algorithms, which reconcile many possible recipes with various inputs, hence with many possible outputs.[52] Due to these learning capacities we cannot passively observe a disclosed code, because what we would see is "a mysterious alchemy in which each individual step might be comprehensible, but any 'explanation' of why the code does what it does requires understanding how it evolved and what 'experiences' it had along the way."[53] Something more than simple observation is required of researchers and social activists seeking to reveal the heart and bones of the codes underlying learning algorithms.

B. *Algorithmic Enforcement on Private Grounds*

The online sphere, in which behavior is inherently mediated by code, has become a prominent site for algorithmic enforcement by private actors.[54] Under the laissez-faire approach to the internet,[55] private, online intermediaries—such as search engines, websites, and social networks—have acquired an important role in managing online behavior and enforcing internet users' rights. They offer a natural point of control for monitoring, filtering, blocking, and disabling access to content, which makes them ideal partners for performing civil and criminal enforcement.[56] Inevitably, these intermediaries often use robots to handle

49. *Id.* (emphasis omitted).

50. *Id.* (emphasis added).

51. *Id.*

52. *See id.*

53. *Id.*

54. *See* Lee, *supra* note 11, at 1035; John Naughton, *How Algorithms Secretly Shape the Way We Behave*, GUARDIAN (Dec. 15, 2012, 7:05 PM), https://www.theguardian.com/technology/2012/dec/16/networker-algorithms-john-naughton.

55. *See generally* WILLIAM J. CLINTON & ALBERT GORE, JR., THE WHITE HOUSE, A FRAMEWORK FOR GLOBAL ELECTRONIC COMMERCE 4 (1997) (proposing that the private sector take the lead on expanding the use of the internet and the government avoid restrictions on electronic commerce).

56. An extensive amount of scholarship has focused on the role of access providers, hosting facilities, search engines, social networks, and application providers as gatekeepers. *See, e.g.*, JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 68 (2006); Patricia Sánchez Abril, *Private Ordering: A Contractual Approach to Online Interpersonal Privacy*, 45 WAKE FOREST L. REV. 689, 721 (2010); Annemarie Bridy, *Graduated*

Electronic copy available at: https://ssrn.com/abstract=2741513

the immense traffic of online content.[57] Examples include the use of different technologies to ensure online content complies with the laws of security,[58] privacy,[59] and intellectual property.[60]

When online intermediaries perform public functions meant to serve the public at large under formal or informal delegation of power from the government, they effectively function like private administrative agencies.[61] That is the case, for instance, with online copyright enforcement pursuant to the Digital Millennium Copyright Act (DMCA).[62] Another example is online enforcement of the Right to Be Forgotten under the recent ruling of the Court of Justice of the European Union (CJEU) in *Google Spain SL v. Agencia Española de Protección de Datos*.[63] As Professor Edward Lee exemplifies:

> Google's role as a private administrative agency is manifest in the variety of public functions it is serving in enforcing the right to be forgotten. It is not surprising that Google describes its own role in classic administrative agency terms: "We had to create an administrative system to intake the requests and then act on them." To put it pithily: Google looks like an agency, talks like an agency, and acts like an agency.[64]

Nevertheless, even though unaccountable law enforcement may lead to manipulation and abuse of power, create new barriers to open

---

Response and the Turn to Private Ordering in Online Copyright Enforcement, 89 OR. L. REV. 81, 84 (2010); Stacey L. Dogan, *Trademark Remedies and Online Intermediaries*, 14 LEWIS & CLARK L. REV. 467, 468–69 (2010); Mark MacCarthy, *What Payment Intermediaries Are Doing About Online Liability and Why It Matters*, 25 BERKELEY TECH. L.J. 1037, 1038 (2010); Ronald J. Mann & Seth R. Belzley, *The Promise of Internet Intermediary Liability*, 47 WM. & MARY L. REV. 239, 254 (2005); Joel R. Reidenberg, *States and Internet Enforcement*, 1 U. OTTAWA L. & TECH. J. 213, 216 (2003–2004); Jonathan Zittrain, *A History of Online Gatekeeping*, 19 HARV. J.L. & TECH. 253, 253–54 (2006).

57. Urban et al., *supra* note 32, at 8.

58. *See* Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT Act) of 2001, Pub. L. No. 107-56, 115 Stat. 272, 290–91 (codified as amended in scattered sections of the U.S. Code) (facilitating government access to customer data held by service providers).

59. *See* 18 U.S.C. § 2702(b)(8) (2012) (granting service providers immunity from damages if, in the case of an emergency, they disclose information to the government about their clients' communications).

60. *See* 17 U.S.C. § 1201 (2012).

61. Lee, *supra* note 11, at 1049.

62. Pub. L. No. 105-304, 112 Stat. 2860 (1998) (codified as amended in scattered sections of the U.S. Code).

63. Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (June 25, 2013), http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&doclang=EN.

64. Lee, *supra* note 11, at 1069–70 (footnote omitted).

Electronic copy available at: https://ssrn.com/abstract=2741513

competition and market innovation, and challenge civil rights,[65] none of the mandatory transparency rules that apply to administrative agencies govern private online intermediaries.[66] For instance, online intermediaries are not required to comply with the Administrative Procedure Act's (APA's)[67] "notice and comment" procedure,[68] or to make all their records available upon public request, as set by the Freedom of Information Act of 1966.[69] As private, profit-maximizing entities, online intermediaries generally have the freedom to manage the content they distribute.[70] Normally, interfering with their internal

65. *See* Jody Freeman, *Private Parties, Public Functions and the New Administrative Law*, 52 ADMIN. L. REV. 813, 818–19 (2000) (acknowledging the potential dangers to democratic accountability that private actors pose in mixed administration); *see also* Lee, *supra* note 11, at 1073–78 (counting several accountability drawbacks arising from giving Google the primary responsibility of deciding the contours of the recently recognized right to be forgotten: private anonymous employees that do not reflect users' diversity, possible bias of employees in favor of access to information, minimal due process afforded to affected users, and mistaken legal determinations); Tal Z. Zarsky, *Social Justice, Social Norms and the Governance of Social Media*, 35 PACE L. REV. 154, 156 (2014) (arguing that "[t]he notion that a small group of managers . . . unilaterally set the rules regulating the social discourse is daunting" and may impact users' core rights, including their "ability to engage in free speech or invoke privacy").

66. *See* Perel & Elkin-Koren, *supra* note 12, at 485–86.

67. Pub. L. No. 79-404, 60 Stat. 237 (1946) (codified as amended in scattered sections of 5 U.S.C. (2012)).

68. *See* 5 U.S.C. § 553(b)–(c). The notice shall include "(1) a statement of the time, place, and nature of public rule making proceedings; (2) reference to the legal authority under which the rule is proposed; and (3) either the terms or substance of the proposed rule or a description of the subjects and issues involved." *Id.* § 553(b).

69. *See* Pub. L. No. 93-502, 88 Stat. 1561 (1974) (codified as amended at 5 U.S.C. § 552).

70. For instance, it is stated in Facebook's Statement of Rights and Responsibilities that "[w]e can remove any content or information you post on Facebook if we believe that it violates this Statement or our policies." *Statement of Rights and Responsibilities*, FACEBOOK https://www.facebook.com/legal/terms (last updated Jan. 30, 2015). YouTube states in section 7.8 of its Terms of Service:

> On becoming aware of any potential violation of these Terms, YouTube reserves the right (but shall have no obligation) to decide whether Content complies with the content requirements set out in these Terms and may remove such Content and/or terminate a User's access for uploading Content which is in violation of these Terms at any time, without prior notice and at its sole discretion.

Terms of Service, YOUTUBE, https://www.youtube.com/static?gl=GB&template=terms (last visited Oct. 8, 2016).

Nevertheless, there is some external intervention in the way internet service providers operate. Under the Open Internet Transparency Rule, they are required to disclose information about "network management practices, performance, and commercial terms of service." *Open Internet Transparency Rule*, FCC, https://www.fcc.gov/guides/open-internet-transparency-rule (last updated Jan. 17, 2017).

Electronic copy available at: https://ssrn.com/abstract=2741513

content-management practices seems as objectionable as meddling with the editorial discretion of publishers of the daily news or the media.[71]

Unfortunately, we cannot simply cure this lack in accountability by demanding full transparency from online intermediaries, especially not if they employ algorithms to regulate online content. First, when online intermediaries use robots rather than humans to fulfill law enforcement tasks, the code they develop and employ is proprietary and thus usually protected under trade secrecy law. In the famous legal battle between Viacom and YouTube, the judge refused to force YouTube to provide Viacom with the computer source code which controls both YouTube.com's search function and Google's internet search tool "Google.com."[72] The court explained that "[t]he search code is the product of over a thousand person-years of work" and "[t]here is no dispute that its secrecy is of enormous commercial value. Someone with access to it could readily perceive its basic design principles, and cause catastrophic competitive harm to Google by sharing them with others who might create their own programs without making the same investment."[73] Although the production and examination of the source code was necessary for Viacom to review YouTube's search algorithm and determine how it handles online copyright infringement, the court denied Viacom's motion to compel YouTube to produce its source code and consequently lose its trade secret.[74]

> The Rule applies to service descriptions, including, for example, expected and actual broadband speed and latency. The Rule also applies to pricing, including monthly prices, usage-based fees, and any other additional fees that consumers may be charged. Additionally, it covers providers' network management practices, such as congestion management practices and the types of traffic subject to those practices.

Id.

71. *See* Associated Press v. United States, 326 U.S. 1, 20 & n.18 (1945) (holding that antitrust law cannot "compel [the Associated Press] or its members to permit publication of anything which their 'reason' tells them should not be published").

72. Viacom Int'l Inc. v. YouTube Inc., 253 F.R.D. 256, 259–60 (S.D.N.Y 2008). Earlier cases invoked trade secrets in Google's ranking algorithm. *See* Kinderstart.com LLC v. Google, Inc., No. C 06-2057 JF (RS), 2006 WL 3246596, at *1, *2 (N.D. Cal. July 13, 2006) (granting Google's motion to dismiss); Search King, Inc. v. Google Tech., Inc., No. CIV-02-1457-M, 2003 WL 21464568, at *3–4 (W.D. Okla. May 27, 2003) (holding that website rankings are protected opinions); Michael J. Madison, *Open Secrects*, *in* THE LAW AND THEORY OF TRADE SECRECY 222, 241 (Rochelle C. Dreyfuss & Katherine J. Strandburg eds., 2011).

73. *Viacom Int'l*, 253 F.R.D. at 259.

74. *Id.* at 260.

Electronic copy available at: https://ssrn.com/abstract=2741513

Secondly, transparency might be futile in this context because encouraging online intermediaries to be more transparent about their law enforcement practices cannot ensure that the public will actually receive better access to meaningful information. As long as online intermediaries stay in the private sphere, disclosure remains entirely voluntary, and the scope of disclosures is left largely unregulated.[75] Because online intermediaries are free to determine what specific information to disclose in accordance with their private, financial interests, providing their voluntary disclosures with determinative weight is like expecting a guard to objectively review its own guardianship. There is a reasonable possibility that the specific information disclosed by such a guard will be incomplete, misleading, or even biased. Therefore, enhanced transparency on the part of private, profit-maximizing platforms seems inherently misplaced, hence generally ineffective.

C. *Robustness or "Too Much Information"*

It is commonly argued that the advent of the internet created unprecedented opportunities for "accessing, sharing, and processing regulatory information."[76] Today, with a click of button we can easily retrieve, access, follow, print, and share information, sparing us the hustle of physically approaching governmental facilities, waiting endlessly in line, submitting paper requests for specific information, and enduring annoying delays in the provisions of the requested information. Yet as promising as this shift may seem, it also introduces a serious problem of magnitude.[77] That is, the ease with which information can be retrieved nowadays, combined with its growing online intensity, opens the floodgates to volumes of data that cannot be read, understood, assimilated

75. Examples in the context of online copyright enforcement include the Chilling Effects Project as well as various transparency reports by different intermediaries. *E.g.*, REDDITSTATIC, REDDIT TRANSPARENCY REPORT (2015), https://www.redditstatic.com/transparency/2014.pdf; *Medium's Transparency Report (2014)*, MEDIUM, https://medium.com/transparency-report/mediums-transparency-report-438fe06936ff (last visited Oct. 8, 2016); *Transparency Report*, GOOGLE [hereinafter *Google Transparency Report*], http://www.google.com/transparencyreport/removals/copyright/ (last visited Oct. 8, 2016); *Transparency Report*, MAPBOX, https://www.mapbox.com/transparency-report/ (last visited Oct. 8, 2016); *Transparency Report*, TWITTER, https://transparency.twitter.com/ (last visited Oct. 8, 2016); *Transparency Report*, WIKIMEDIA FOUND., https://transparency.wikimedia.org/ (last visited Oct. 8, 2016); *Transparency Report*, WORDPRESS, http://transparency.automattic.com/ (last visited Oct. 8, 2016).

76. Shkabatur, *supra* note 35, at 80.

77. *See* NIVA ELKIN-KOREN & ELI M. SALZBERGER, LAW, ECONOMICS AND CYBERSPACE 70, 94–96 (2004) (arguing that while the costs of retrieving information in cyberspace may go lower, the cognitive barriers on individual choice are likely to become stronger); Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U. PA. L. REV. 647, 686 (2011) (explaining the "quantity problem" of mandated disclosure).

Electronic copy available at: https://ssrn.com/abstract=2741513

or analyzed.[78] As Professors Omri Ben-Shahar and Carl E. Schneider have recently shown, the amount of information provided by disclosures tends to paralyze people.[79] For example, only one or two in a thousand consumers actually scrolls the terms of use provided by online service providers, such as iTunes, before clicking "I agree,"[80] indicating how overwhelming the platforms' respective disclosures are.[81]

Of course, countless other online service providers offer similar disclosures of unintelligible, often not particularly readable material, further undermining the principal objective of transparency: to inform online users about how their online conduct is regulated. Professors Ben-Shahar and Schneider named this aspect of the information-magnitude problem "the 'accumulation' problem," explaining that "each disclosure competes for [users'] time and attention with other disclosures, with their investigations into unmandated knowledge, and with everything they do besides collecting information and making decisions (like working, playing, and living with their families)."[82] Indeed, "[e]ven if disclosees wanted to read all the disclosures relevant to their decisions, they could not do so proficiently, and practically they could not do so at all."[83]

In fact, analyzing this overflow of disclosed data in itself requires algorithmic processing that is capable of turning the data into meaningful information.[84] Yet this creates a vicious cycle: More transparency only

78. *See* Howard Latin, *"Good" Warnings, Bad Products, and Cognitive Limitations*, 41 UCLA L. REV. 1193, 1211–15 (1994) (examining the social science literature on information overload).

79. OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE 74–78 (2014) (explaining the many reasons people ignore mandatory disclosures, including, for example: believing it is irrelevant, can be ignored, will not be understood anyway, and is too boring to read).

80. *See* Yannis Bakos et al., *Does Anyone Read the Fine Print? Consumer Attention to Standard Form Contracts* 2 (N.Y. Univ. Law & Econ., Working Paper No. 195, 2014), http://lsr.nellco.org/nyu_lewp/195/.

81. That one-in-a-thousand "reader" spends a median time of twenty-nine seconds skimming a word document of around 2,000 words, which means that the actual readership is effectively zero, considering the average reading rate of 250 to 300 words per minute. *Id.* at 2, 22.

82. Ben-Shahar & Schneider, *supra* note 77, at 689.

83. *Id.* at 690.

84. For instance, Joel R. Reidenberg, Jaspreet Bhatia and Travis D. Breauk recently addressed the technology of Natural Language Processing (NLP) that is capable of identifying and measuring ambiguity in website policies, while providing companies with a useful mechanism to improve the drafting of their policies. *See* Joel R. Reidenberg et al., *Ambiguity in Privacy Policies and the Impact of Regulation*, 45 J. LEGAL STUDIES, S163, S163 (2016). Nevertheless, how can one examine whether the NLP algorithm is fair and trustworthy? While it arguably mitigates the magnitude problem of disclosures, it reinforces the inherent problems of opacity and machine learning discussed in Section I.A.

Electronic copy available at: https://ssrn.com/abstract=2741513

strengthens users' dependence on algorithms, which further increases the need to ensure adequate accountability of the algorithms themselves.[85]

Voluntary disclosures of online intermediaries engaging in copyright enforcement afford a classic example of the information-magnitude problem. Indeed, as this Essay explains further in Part II, copyright enforcement through online mechanisms run by algorithms has become robust.[86] Copyright owners prefer to vindicate their rights by resorting to online resolution systems, instead of going through the hustle of filing an expensive and time-consuming suit for copyright infringement.[87] To be even more efficient, many copyright owners use robots to search the web for infringing activity, submitting huge amounts of automatic removal requests simultaneously to all platforms identified as containing allegedly infringing material.[88] As a result, voluntary reports of complaints received by major online platforms are inevitably overwhelming.

Google's Transparency Report, for instance, publishes tens of millions of copyright removal requests received for Google *Search* each month.[89] This, of course, does not include removal requests received for Google *Image*, or removal requests received by other prominent platforms such as Facebook, YouTube, or Twitter. Thus, to rely on voluntary reports of online intermediaries in order to draw intelligible insights about their copyright enforcement practices, it is necessary not only to have a useful methodology to retrieve protected or undisclosed information held by private entities,[90] but also to address and analyze the available information. When disclosures are too many and involve too much information, as happens with robust online copyright enforcement systems, merely having access to relevant information is simply not enough to generate adequate accountability. Proper tools to accumulate and interact with the information become essential to turning the data into meaningful information and making transparency a useful tool for enhancing accountability.

---

85. *See supra* Section I.A.

86. *See infra* Section II.B.

87. *See* H.R. REP NO. 105-796, at 72 (1998) (Conf. Rep.) ("Title II preserves strong incentives for service providers and copyright owners to cooperate to detect and deal with copyright infringements that take place in the digital networked environment. At the same time, it provides greater certainty to service providers concerning their legal exposure for infringements that may occur in the course of their activities."); David Nimmer, *Appreciating Legislative History: The Sweet and Sour Spots of the DMCA's Commentary*, 23 CARDOZO L. REV. 909, 917–18 (2002).

88. Urban et al., *supra* note 32, at 31–32.

89. *Google Transparency Report*, *supra* note 75.

90. *See supra* Section I.B.

Electronic copy available at: https://ssrn.com/abstract=2741513

D. *Legal Discretion*

Finally, employing algorithms to engage in discretional legal analysis is another reason that transparency is an insufficient guardian of algorithmic enforcement. With algorithms that detect strict liability, transparency about the algorithmic outcome provides full, or relatively full, information about the algorithmic process. For example, the outcome of an automatically issued speeding ticket means that the targeted car was automatically photographed by a police radar for exceeding the statutory speed limit. Because the algorithm which underlies the radar does not consider external circumstances that may affect a given driving speed (such as the driver's attempt to bring an injured passenger to the nearest ER), the meaning of a speeding ticket is fairly straightforward. However, unlike determining strict liability rules, the implementation of flexible standards inherently depends on the consideration and weight of qualitative factors made on a case-by-case basis. Merely observing the outcome of applying a discretional legal standard to a given set of circumstances teaches very little about the underlying process of weighing and assessing the relevant factors.

Consider, for example, algorithmic copyright enforcement. Many of the most serious issues in copyright law involve discretion, including determining the degree of "originality" required to establish copyrightability;[91] deciding what amounts to "substantial similarity" to establish infringement;[92] or considering what constitutes "permissible use" under fair use.[93] Before algorithms can implement these qualitative doctrines, they must first be translated into "codish" thresholds, a process that in itself may result in unintentional alterations of settled doctrines. Accordingly, restriction of content due to copyright infringement yields nothing about how the algorithm has effectively applied the four-factor test of fair use.[94] Has it considered the effect of the allegedly infringing use on the potential market for the protected work? Has it identified the nature of use? Its purpose? What if the algorithm has only considered what portion of the protected work was taken, while ignoring the other qualitative factors—is it possible to unveil such a distorted application of the fair use doctrine by simply observing a given outcome of content restriction? The answer is "maybe." Unfortunately, it is unclear how online mechanisms of algorithmic copyright enforcement exercise their power.

---

91. *See* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 345 (1991).

92. *See* Ideal Toy Corp. v. Fab-Lu Ltd., 266 F. Supp. 755, 756 (S.D.N.Y. 1965), *aff'd*, 360 F.2d 1021 (2d Cir. 1966).

93. *See* Cambridge Univ. Press v. Patton, 769 F.3d 1232, 1282 (11th Cir. 2014).

94. *See infra* note 155 and accompanying text.

Electronic copy available at: https://ssrn.com/abstract=2741513

But if it were possible to *interact* with algorithmic systems of online copyright enforcement there might be an opportunity to learn more about their practices.[95] For instance, if it were possible to submit artificial content containing similar portions of protected materials, but deviating in the purpose of use, it would be possible to determine whether the tested system of algorithmic copyright enforcement actually considers the nature of the allegedly infringing use before restricting content.[96] Similarly, if it were possible to submit content differentiating in the quantitative portion of protected material taken, it might reveal the underlying algorithm's quantitative threshold of copyright infringement.

To conclude, algorithmic enforcement by private entities raises serious challenges to the notion of transparency as the principal guardian of decision makers' accountability. The resort to complex learning codes that are employed by private profit-maximizing entities to implement flexible legal standards introduces a new form of black box governance, which cannot be easily reviewed. Hence the use of more active accountability-enhancing tools must be deemed to increase the ability of the general public to interact with the complex machines that regulate their behavior and reveal their internal operations. The following Part describes the methodology of black box tinkering and demonstrates its benefits in advancing social activism, using the test case of algorithmic copyright enforcement by online intermediaries.

II. BLACK BOX TINKERING

Computer scientist Edward Felten has defined the term "Freedom to Tinker" as "[people's] freedom to understand, discuss, repair, and modify the technological devices [they] own."[97] According to Professor Pamela Samuelson, people tinker with technologies

> to have fun, to be playful, to learn how things work, to discern their flaws or vulnerabilities, to build their skills, to become more actualized, to tailor the artifacts to serve one's specific needs or functions, to repair or make improvements to the artifacts, to adapt them to new purposes, and occasionally, to be destructive.[98]

95. *See infra* Subsection II.B.4.

96. *See infra* note 155 and accompanying text.

97. Ed Felten, *The New Freedom to Tinker Movement*, FREEDOM TO TINKER (Mar. 21, 2013), https://freedom-to-tinker.com/blog/felten/the-new-freedom-to-tinker-movement.

98. Pamela Samuelson, *Freedom to Tinker* 1–2 (U.C. Berkeley, Pub. Law & Legal Theory Research Paper Series, Paper No. 2605195, 2015), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2605195.

Electronic copy available at: https://ssrn.com/abstract=2741513

Intellectual property scholars tend to associate freedom to tinker with the freedom to innovate,[99] and Professor Samuelson adds that freedom to tinker is important also to encourage "freedom of thought, study, inquiry, self-expression, [and] diffusion of knowledge," and to "foster[] privacy, autonomy, human flourishing, and skills building interests of tinkerers."[100] Obviously, as more and more aspects of our daily conduct are mediated through technology, freedom to tinker becomes crucial to help us define our relationship with the governing technology.[101]

For instance, a black box tinkering experiment recently conducted in China by Professors Gary King, Jennifer Pan, and Margaret E. Roberts delivered meaningful insights about political censorship in social media sites.[102] The research team created accounts on numerous social media sites across China, while randomly submitting a large number of unique social media posts written by the research team for the purpose of the study.[103] Afterwards, from a worldwide network of computers, the research team observed which posts were censored and which were not.[104] By actively tinkering with the black box system of online censorship in China, the researchers were able to obtain very interesting results that could not have been done through alternative observational methodologies. For instance, they found that automated review affects a remarkably large portion of the social media landscape in China, with more than 60% of the sites examined employing automatic review of at least some social media submissions, practically changing the default from "publish first, censor later," to "review first, maybe publish later."[105] This surprising finding must surely encourage further research into which keywords provoke automatic action by the government, how automated review works, and what impact this process ultimately makes on the content of speech that is blocked and on that which can be consumed by the Chinese people.[106]

Beyond facilitating human interaction with the black box systems surrounding them, this Essay argues that freedom to tinker may further facilitate social activism, creating a policy lever for checks and balances

99. *See, e.g.*, Andrew W. Torrance & Eric von Hippel, *The Right to Innovate*, 2015 MICH. ST. L. REV. 793, 807.

100. Samuelson, *supra* note 98, at 21.

101. *See* Felten, *supra* note 97.

102. *See* Gary King et al., *Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation*, 345 SCIENCE 891, 899 (2014).

103. *Id.* at 894.

104. *Id*. at 895.

105. *Id.* at 895, 899.

106. *See* Jedidiah R. Crandall et al., *Chat Program Censorship and Surveillance in China: Tracking TOM-Skype and Sina UC*, FIRST MONDAY (July 1, 2013), http://firstmonday.org/ojs/index.php/fm/article/view/4628/3727.

Electronic copy available at: https://ssrn.com/abstract=2741513

of the hidden practices of non-transparent algorithms. Specifically, it is an important tool for proactively checking the credibility, fairness, and trustworthiness of algorithms that cannot be adequately reviewed through traditional means of transparency.[107] Black box tinkering can encourage the public to exercise reasonable judgment and demand that algorithmic systems comply with public interests such as due process, equal protection, and freedom of expression.[108] Presumably, thanks to Professors King, Pan and Roberts' active attempt to challenge the online censorship system in China, the public now knows that a large proportion of their online submissions are targeted automatically, before going through human review. With the aid of sequential research regarding which specific keywords are likely to ignite automatic action by the government, participants on the Chinese social media scene may be encouraged to use their wording in a way that will allow them to bypass automatic treatment. They may further elect to limit their online activity to social media sites that do not apply automatic review, thereby ensuring that their posts are at least reviewed manually, according to China's censorship agenda, and are not arbitrarily signaled and blocked by automated machines.

In the following discussion this Essay further explores the benefits of black box tinkering, demonstrating their application in the context of algorithmic copyright enforcement. It provides a brief background of algorithmic copyright enforcement then describes a recent study applying black box tinkering methodology to investigate algorithmic copyright enforcement practices by online intermediaries. After discussing the study's findings, the Essay explains their broader implications for algorithmic copyright enforcement: further strengthening the benefits of black box tinkering as a valuable oversight tool.

A. *The Benefits of Black Box Tinkering*

To appreciate the virtues of black box tinkering, it is helpful to compare it to alternative methodologies for learning about the actual practices of hidden algorithms. One such methodology is observational studies. For instance, to learn about political censorship in China, Professors King, Pan and Roberts' tinkering approach involved randomly submitting different types of texts and recordings.[109] Instead, it is possible to conduct quantitative studies that passively analyze large quantities of social media posts from different websites across China.[110] One recent

107. *See infra* Section II.A.

108. *See* Perel & Elkin-Koren, *supra* note 12, at 519–20, 531–32.

109. *See* King et al., *supra* note 102, at 895–96.

110. *See generally* Gary King et al., *How Censorship in China Allows Government Criticism but Silences Collective Expression*, 107 AM. POL. SCI. REV. 326, 326 (2013) (discussing a large-

Electronic copy available at: https://ssrn.com/abstract=2741513

observational study created a means of analyzing the content of millions of social media posts from all over China before the Chinese government was able to censor the objectionable content.[111] Using modern computer-assisted text analytic methods, the study compared the substantive content of censored and non-censored posts over time in eighty-five different topic areas.[112] Interestingly, posts with scathing critique of the state or its leaders were no more likely to be censored. In fact, the censorship program primarily focused on content that could spur social action.[113]

Although these findings are indisputably inspiring, their limited breadth demonstrates two major downsides of counting solely on observational studies to elicit full accountability over algorithmic enforcement by private actors. First, conclusions derived from observational studies can only reflect publicly available data—in this example social media posts available online. Observational studies that rely on qualitative methodologies of this sort, especially when analyzing the practices of private, profit-maximizing actors such as online platforms, are confined to investigating the data that intermediaries voluntarily make public.[114] Yet since platforms may purportedly omit information that interferes with their economic interests or otherwise violates their business obligations to their partners, voluntarily disclosed data may be distorted by unavoidable bias. Hence, observational studies may occasionally provide inadequate checks, especially where the review process is contingent on data that the subject of review voluntarily chooses to share.

Second, observational studies that rely on qualitative methodologies are ex post by definition. They ultimately examine after-the-fact by-products of a given phenomenon, but not the phenomenon itself as it occurs in real time. For instance, the qualitative investigation of political censorship in China discussed above compared the substantive content of posts that were not blocked ex ante from publication by the automated review process, but were nonetheless censored (removed from the

---

scale, multiple-source analysis of the outcome of an extensive effort to sensor human expression in China); Tao Zhu et al., *The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions*, 22 USENIX SECURITY SYMP. 227 (2013) (analyzing how fast and how comprehensively posts are deleted from Chinese microblogging sites due to internal censorship); David Bamman et al., *Censorship and Deletion Practices in Chinese Social Media*, FIRST MONDAY (Mar. 5, 2012), http://journals.uic.edu/ojs/index.php/fm/article/view/3943/3169 (discussing the first large-scale statistical analysis of political content censorship in Chinese social media).

111. King et al., *supra* note 110, at 326.

112. *Id.*

113. *Id.*

114. *See supra* note 110 and accompanying text.

Electronic copy available at: https://ssrn.com/abstract=2741513

internet), with the content of posts not censored.[115] Because it applied an ex post, observational methodology, this study could not address content that was automatically and instantaneously filtered and slotted for subsequent manual review by an ex ante, automated content review process.[116] Effectively, the study primarily analyzed the by-products of political censorship in China—the actual removed submissions—while affording less attention to the actual act of censorship, as it transpired in real time.

The proactive methodology of black box tinkering can overcome these two major limitations of observational studies. First, it can defuse the possibility of bias arising from relying on data voluntarily disclosed by private, profit-maximizing platforms, for black box tinkering researchers enjoy the benefit of independently extracting their own random database. So unlike observational researchers engaging in qualitative analysis, black box tinkering researchers do not depend on platforms' disclosure and are not controlled by an earlier stage of possible data suppression.[117] As a result, they can avoid both the magnitude and the possible partialness of voluntary transparency reports. Second, black box tinkering allows researchers to elect what exact type of conduct—ex post or ex ante—to investigate. Black box tinkering researchers can therefore go beyond observing the ex post by-products of a given phenomenon, and actively trigger the algorithm underlying the phenomenon to extract its blueprints. Therefore, it furnishes a useful tool to examine the operation of hidden algorithms and paint a better picture of the system studied.

B. *Testing the Black Box Tinkering Methodology*

This Section further demonstrates the benefits of the black box tinkering methodology using an example from a recent case study involving algorithmic copyright enforcement by online intermediaries.

1. The Case Study of Algorithmic Copyright Enforcement by Online Intermediaries

Copyright law has been at the forefront of digital law enforcement since the early 1990s. The ease of digital copying and mass distribution gave rise to digital locks, Digital Rights Management (DRM) systems, and Technological Protection Measures (TPM), which enable right-holders technically to prevent unauthorized access to and use of their

115. King et al., *supra* note 102, at 892.

116. *Id.*

117. Such as automatic content restrictions that block researchers' access to particularly interesting data.

Electronic copy available at: https://ssrn.com/abstract=2741513

copyrighted works.[118] Yet confronted by the threats of dispersed mass piracy, right-holders increased their pressure on Online Service Providers (OSPs) to participate actively in online copyright enforcement.[119] OSPs, for their part, saw the free flow of information as essential to their business models and attempted to avoid the cost of online enforcement.[120] This battle between right-holders and OSPs shaped the intermediary safe harbor regime under the Digital Millennium Copyright Act, which helped copyright owners ensure rapid removal of allegedly infringing material from the internet, while guaranteeing compliant OSPs a safe harbor from liability for internet users' acts of copyright infringement.[121]

The Notice and Takedown (N&TD) procedure established by the DMCA requires OSPs to respond "expeditiously" to notices of infringement by removing or disabling access to allegedly infringing material when certain conditions are met.[122] A hosting service (website, social network) is further required to take "reasonable steps promptly to notify the subscriber that it has removed or disabled access to the material"[123] and promptly to forward any counter notices from alleged infringers back to the original complainant.[124] If after ten to fourteen days following receipt of the counter notice the complainant does not notify the OSP that she has filed a lawsuit, the OSP must reinstate the contested material.[125]

---

118. For a discussion of these technologies see, for example, Pamela Samuelson, *Intellectual Property and the Digital Economy: Why the Anti-Circumvention Regulations Need to Be Revised*, 14 BERKELEY TECH. L.J. 519 (1999). *See also* Timothy K. Armstrong, *Digital Rights Management and the Process of Fair Use*, 20 HARV. J.L. & TECH. 49, 50–51 (2006) (discussing a multitude of current and proposed DRM systems).

119. Niva Elkin-Koren, *After Twenty Years: Revisiting Copyright Liability of Online Intermediaries*, *in* THE EVOLUTION AND EQUILIBRIUM OF COPYRIGHT IN THE DIGITAL AGE 29, 29, 31–32 (2014) ("Digital networks have led to an 'enforcement failure' in copyright-related industries, turning online intermediaries into key players in enforcement efforts.").

120. *Id.* at 31.

121. To maintain immunity from monetary liability for material that is transmitted over networks, cached on a server, or linked to or stored at the direction of a user, OSPs were required to adopt and implement certain policies. In particular, OSPs must comply with two preliminary policies. First, they must adopt and reasonably implement a policy to terminate the accounts of repeat infringers and must notify users of this plan. Second, they must also accommodate "standard technical measures" used by copyright owners to identify infringing material. *See* 17 U.S.C. § 512(a)–(d), (i) (2012).

122. *Id.* § 512(b)(2)(E)(i)–(ii), (c)(1)(C).

123. *Id.* § 512(g)(2)(A).

124. *Id.* § 512(g)(2)(B). A counter notification must include the following: (A) a physical or electronic signature; (B) identification of the material removed and its former location; (C) a statement under penalty of perjury that the user has a good faith belief the material was mistakenly removed; (D) the user's name, address, and phone number; and consent to the jurisdiction of Federal District Court. *Id.* § 512(g)(3).

125. CHEN ET AL., *supra* note 32, at 16–17. Search engines, on the other hand, are not required to notify the alleged infringer of removal because they are not expected to have any service relationship with the alleged infringer. 17 U.S.C. § 512(d).

Electronic copy available at: https://ssrn.com/abstract=2741513

To maintain its immunity under the N&TD regime, an OSP cannot have actual knowledge that infringing content is on its system or be "aware of facts or circumstances from which infringing activity is apparent."[126] Moreover, it should not receive a direct financial benefit from any infringing activity it has the right and ability to control.[127] Finally, the DMCA further encourages compliance with N&TD by exempting OSPs from liability for mistaken yet good faith removal of material.[128]

Arguably, copyright enforcement by online intermediaries pursuant to the DMCA's N&TD framework offers an efficient alternative to the cumbersome, often impracticable traditional enforcement of copyrights through the legal system, which barely keeps up with the accelerated pace of technological change. The legal system is often understaffed, slow to act, and costly for litigants and for society[129] compared with the low cost, instant, scalable, and robust system of online copyright enforcement. This helps explain why much of modern online copyright enforcement is embedded in the system design of online intermediaries, using algorithms to remove allegedly infringing content upon notice of copyright infringement and to monitor, filter, block, and disable access to material that is automatically flagged as infringing.[130]

In fact, recent studies prove that prominent OSPs, facing a flood of robo-takedown notices sent automatically by right-holders, substitute human review of the vast majority of these notices with their own privately designed automated systems.[131] Relying on smart algorithms to enforce the rights of copyright owners, these automated systems effectively manage the distribution of online content.[132] But when is algorithmic enforcement employed? Can we judge whether automated systems comply with the rule of law they enforce (i.e., the DMCA)? In particular, do we know how they quantify the flexible standards of

126. § 512(c)(1)(A). If OSPs later become aware of such content, they must expeditiously remove it from their system. *Id.*

127. *Id.* § 512(c)(1)(B).

128. *Id.* § 512(g)(1). Intermediaries that fail to act in good faith may lose safe harbor and may be required to pay damages to content providers whose material was unlawfully removed under the intermediaries' stated terms of use.

129. For similar arguments in relation to risk management, see Bamberger, *supra* note 9, at 685.

130. Perel & Elkin-Koren, *supra* note 12, at 484–85.

131. Urban et al., *supra* note 32, at 10; *see also* Daniel Seng, *The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices*, 18 VA. J.L. & TECH. 369, 444, 460–61 (2014) (providing charts analyzing data of takedown notices).

132. Perel & Elkin-Koren, *supra* note 12, at 476, 479.

Electronic copy available at: https://ssrn.com/abstract=2741513

copyright law, such as "substantial similarity,"[133] or fair use?[134] Are these automated systems subject to abuse by interested parties that wish to influence public discourse and silence legitimate speech?[135]

To begin answering these cardinal questions, this Essay seeks to shed some light on the hidden practices of online platforms in enforcing the rights of copyright holders. We used the methodology of black box tinkering as our flashlight.

2. Study Description

The study, conducted in 2013 by research students at the Haifa Center for Law and Technology, was conducted outside the United States, in Israel, where there is no clear statutory framework governing the N&TD regime. As under U.S. law, however, online intermediaries might be subject to contributory liability for infringing materials posted by the subscribers to their service, if they fail to remove the materials on receiving a notice.[136] The Israeli N&TD regime provides two major research benefits. The first is technical and relates to size: Since Israel is a very small jurisdiction, researchers can conduct a comprehensive study of all the relevant online platforms, with no practical need to limit their investigation to a representative sample of a large dataset. The second benefit is substantial: The Israeli N&TD procedure lacks the formalities defined by the DMCA.[137] As a result, researchers enjoy greater leeway to tinker with online mechanisms of algorithmic copyright enforcement. Indeed, many of the U.S. legal barriers this Essay discusses in Part III—especially the requirement that complainants verify their copyright ownership under oath, before submitting a notice of infringement[138]—are simply non-existent under the Israeli copyright regime, making it easier to apply black box tinkering for research purposes.

---

133. *See, e.g.*, Steinberg v. Columbia Pictures Indus., 663 F. Supp 706, 711 (S.D.N.Y. 1987); Ideal Toy Corp. v. Fab-Lu Ltd., 266 F. Supp. 755, 756 (S.D.N.Y. 1965), *aff'd*, 360 F.2d 1021 (2d Cir. 1966).

134. Cambridge Univ. Press v. Patton, 769 F.3d 1232, 1238 (11th Cir. 2014).

135. *See* Perel & Elkin-Koren, *supra* note 12, at 482–83, 488.

136. DC (Central District) 567-08-09 *ALIS—Association for the Protection of Cinematic Works v. Rotter.net Ltd.* (2011) (Isr.). Online intermediaries who fail to remove infringing materials of their subscribers upon receiving a notice may face contributory liability. Note that liability for contributory copyright infringement under Israeli case law requires knowledge of the infringing acts. Constructive knowledge would be insufficient for establishing liability. *See* CA 5977/07 *The Hebrew University of Jerusalem v. Shoken Publishing Ltd. and Others* (2011) (Isr.).

137. The following discussion about the legal barriers to enhancing algorithmic accountability explains why conducting such a stimulating experiment in the United States is currently highly problematic. *See infra* Part III.

138. 17 U.S.C. § 512(c)(3)(A) (2012); *see infra* notes 175–77 and accompanying text.

Electronic copy available at: https://ssrn.com/abstract=2741513

Essentially, the study sought to test systematically how hosting websites implement the N&TD policy by examining popular local image-sharing and video-sharing platforms.[139] Accordingly, different types of infringing, non-infringing and fair use materials were uploaded to the hosting facilities, each intended to trace choices made by the black box system throughout its enforcement process.[140]

One set of questions focused on the detection of infringing materials and whether sites undertake any proactive measures to remove materials, or whether they only act on a notice. For instance, to determine whether a hosting facility uses a filter to detect presumably infringing materials, two versions of a video snippet were uploaded to different video-sharing platforms. One was short, taken from an original copyright-protected video with Content ID, the other identical but without Content ID.[141] Comparing how video-sharing sites handled the different snippets allowed determining whether sites were using an automated system to detect and block content ex ante based on a digital signature, or whether they only removed such posts ex post, on receiving a notice. Another type of video snippet, which included non-infringing footage but samples of copyrighted music, allowed identifying which signals were tracked by filters.[142] A non-infringing snippet was also uploaded, to determine whether the system filtered those materials.[143]

Another set of questions focused on the processing of infringement notices. Accordingly, to test whether the removal is automatic on receipt of a notice of copyright infringement or whether hosting platforms implement some measure of discretion before removing allegedly infringing content, the study design included some content intended to stimulate doubt about the submitted takedown notices. An example is a homemade video clip of a toddler dancing to a few bars of copyrighted

139. These platforms were designated by various Israeli forums as being the most popular file-sharing platforms in Israel, a designation that was also confirmed by the second biggest advertising company in Israel.

140. The researchers attempted to upload three types of images to the image-sharing platforms: (1) an infringing image of a known brand with a copyright notice, ©; (2) a non-infringing image; and (3) a non-infringing image with a copyright notice, ©. The researchers also attempted to upload different types of video snippets to the video-sharing platforms: (a) a 2:42-minute infringing video with Content ID (a short snippet of an original, copyright-protected video can trigger an automatic content filtering technology, such as YouTube's Content ID); (b) a similar snippet of the same video, but without Content ID (a snippet of an already-copied video may not be identified by an automatic content filtering technology such as YouTube's Content ID); (c) a non-infringing short video; (d) a fair use homemade video clip; (e) a 19-second non-infringing video with a copyright notice, ©; and (f) a 3:22-minute video of non-infringing photos with a copyright notice, ©, and with an infringing music.

141. *See supra* note 140 (referring to contents (a) & (b)).

142. *See supra* note 140 (referring to content (f)).

143. *See supra* note 140 (referring to content (c)).

Electronic copy available at: https://ssrn.com/abstract=2741513

background music, intended to test whether fair use was considered prior to removal.[144] Also, a clearly non-infringing homemade video with random sounds which contained a visible copyright notice, ©, and an FBI anti-piracy warning was uploaded to identify automatic processing.[145]

Furthermore, in the absence of any statutory framework for verifying rights under Israel's N&TD regime, the study sought to examine whether hosting sites still requested any identifying information from complainants or otherwise attempted to verify their rights and identity. Accordingly, a clearly infringing image of a famous brand, with visible copyright and trademarks notices, was uploaded, and an anonymous user who was clearly not associated with the multinational brand submitted a takedown notice.[146]

Finally, the study tracked whether the tested hosting sites responded to removal notices and how long it took. Further issues regarding the process were examined, such as whether hosting sites notify alleged infringers and complainants about content removals and whether the content becomes accessible following the removal.[147]

The study proceeded in several steps, each systematically recorded by the researchers: First, the researchers submitted different types of content to the examined platforms.[148] When upload was unsuccessful, the researchers assumed that an ex ante mechanism of filtering was used. Second, when upload was successful, the researchers checked periodically whether the content remained online or was otherwise blocked or removed. Third, if the content remained online after seventy-two hours, the researchers sent a notice to the platform complaining it was probably hosting copyright infringing content. Fourth, if the content was removed by the platform after receiving the notice, the researchers reported whether they received a notice of removal. If they were notified about the removal, the researchers reported whether the notification contained information about the removal reason, whether it contained information about the complainant, and whether it provided any dispute opportunities. Fifth, the researchers examined periodically whether the removed content remained offline.

144. *See supra* note 140 (referring to content (d)).

145. *See supra* note 140 (referring to content (e)).

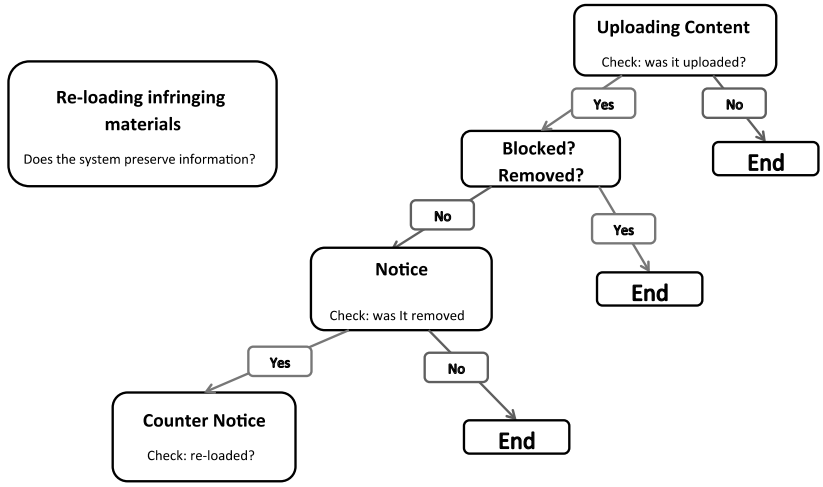
146. *See supra* note 140 (referring to content (1)).

147. The study received the approval of the Israeli Ethical Committee. Moreover, upon the completion of the experiment, all platforms were notified that they had participated in a study in social science testing their online copyright enforcement practices, and best efforts were made to subsequently cause the removal of any infringing content which was successfully uploaded to the platforms.

148. *See supra* note 140.

Electronic copy available at: https://ssrn.com/abstract=2741513

Pilot Flowchart



3. Findings

The findings of the study demonstrate that local hosting platforms in Israel are inconsistent and therefore unpredictable in detecting online infringement and enforcing copyrights. Specifically, 25% of video-sharing platforms and approximately 10% of image-sharing platforms seem to employ a system of ex ante filtering of presumably infringing online content, indicating that online intermediaries occasionally go beyond N&TD, removing content automatically before receiving a complaint about copyright infringement. Furthermore, 50% of video-sharing platforms but only 12.5% of image-sharing platforms removed infringing content after receiving a complaint notice. Absurdly, image-sharing platforms were more responsive to complaint notices addressing *non-infringing* material with one third of the total image-sharing sites unjustifiably removing non-infringing content after receiving a complaint notice. This means that some platforms allow content that is filtered by others; some platforms strictly respond to *any* notice requesting removal of content despite its being clearly non-infringing, while other platforms fail to remove content upon notice of alleged infringement.

Moreover, 75% of video-sharing platforms and 44% of image-sharing platforms required complainants to verify their identity before filing a notice of removal. 50% of video-sharing platforms but only 22% of image-sharing platforms further required complainants to verify that they were the lawful owners of the copyright claimed to be infringed. This suggests that many online mechanisms of algorithmic copyright

Electronic copy available at: https://ssrn.com/abstract=2741513

enforcement generally do very little in terms of minimizing errors and ensuring that interested parties do not abuse the system to silence legitimate speech and over-enforce copyright.[149]

Finally, an important part of managing an appropriate online enforcement procedure for the removal of allegedly infringing content is notifying complainants that their notice of removal has been received, and subsequently notifying alleged infringers of the removal of their content; still, the findings show that image-sharing platforms rank poorly in facilitating an adequate process of dispute resolution, with only one third of image-sharing platforms notifying complainants about receiving their complaints, compared with 75% of video-sharing platforms. Moreover, all video-sharing platforms, but only 11% of image-sharing platforms, had afterwards notified alleged infringers about the removal of their content. This indicates that platforms do not make full efforts to secure due process and allow affected individuals to follow, and promptly respond to, proceedings that manage their online submissions.

4. Lessons on Algorithmic Copyright Enforcement by Online Intermediaries

The black box tinkering study described above offers an invaluable grasp of online copyright enforcement practices on the ground. By actively challenging black box systems of online content adjudication[150] with real world content, the study reveals how diverse and inconsistent these systems actually are. The fact that N&TD in Israel is not statutorily regulated apparently left local platforms with somewhat generous leeway to design their own content-removal policies, often with minimal DMCA-style procedural safeguards, such as the demand that copyright owners verify their rights under oath,[151] or the requirement that platforms provide affected users with removal notification and dispute opportunities.[152] This may mean that a non-regulated regime of N&TD is not necessarily superior to a regulated system, at least not in relation to pursuing procedural justice.

Furthermore, the study successfully extracted substantial evidence about algorithmic errors far beyond currently available anecdotes of

149. *See supra* note 22.

150. *See* Perel & Elkin-Koren, *supra* note 12, at 482–84 (arguing that when online platforms determine the legitimacy of online content based on external legal mandates, such as IP law or the right to be forgotten, their determinations can no longer be viewed as expressions of private content management—rather, they de facto constitute a manifestation of judicial-style content adjudication).

151. *See* 17 U.S.C. § 512(c)(3)(A) (2012).

152. *Id.* § 512(g)(2).

Electronic copy available at: https://ssrn.com/abstract=2741513

erroneous content restrictions.[153] By systematically attempting to upload non-infringing content and recording the responses of platforms, we could easily detect instances of false positives (i.e., removal of non-infringing content). Similarly, the study further provided empirical support for the common proposition that algorithms may reflect a wrong interpretation of the law they enforce,[154] as manifested by the occasional mistaken fair use analysis some of the tested platforms had conducted.[155] Finally, the study surprisingly rebutted our intuitive expectation that small, local platforms would not resort to automatic filtering of allegedly infringing content, but engage in manual, case-by-case examinations.[156]

---

153. See Perel & Elkin-Koren, supra note 12, at 504–07.

154. *See* Bamberger, *supra* note 9, at 675–76.

155. Specifically, when the researchers tried to upload a 48-second homemade video of a child dancing to a protected song by singer Justin Bieber, 25% of video-sharing platforms removed the video, notwithstanding it clearly constituted a fair use. Under 17 U.S.C. § 107, the factors to be considered when determining whether the use made of a work in any particular case is a fair use include—

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.

17 U.S.C. § 107. Applying these factors to the 48-second homemade video used in the study we described clearly suggests that it qualifies as fair use: it only made a *de minimis* use of the protected song, for private, non-commercial purposes that cannot possibly affect the potential market for the original song.

156. However, Urban, Karaganis, and Schofield have recently reached an opposite conclusion in their qualitative study of N&TD practices, stressing that: "[N]otice and takedown continues to operate largely as it has since the DMCA took effect, *without* large-scale notice sending and handling." Urban et al., *supra* note 32, at 10. Many OSPs—"often, but not always, companies outside of the contested music, video, and search areas"—are not pressured by the need to process large numbers of automated notices. *Id.* These "DMCA Classic" OSPs described evaluating and processing takedown requests in essentially the same way as when N&TD was first adopted—typically reviewing individual notices, sometimes very few, by hand. *Id.* at 2. Nevertheless, it is important to note that these conclusions are derived from observational research methodology based on a survey among several cooperating platforms. *See id.* at 26–27. As explained earlier, qualitative observations of this sort may sometimes be misleading. First, they only reflect the practices of cooperative platforms (less than thirty-six in this specific study by Urban, Karaganis, and Schofield). *Id.* at 26. Yet what if the non-surveyed platforms overwhelmingly resort to automatic content review? Second and related, surveys and interviews only reflect what the surveyed object is willing to admit. In that sense, the Urban, Karaganis, and

Electronic copy available at: https://ssrn.com/abstract=2741513

Accordingly, supplementing data from observational studies[157] and anecdotal reports on algorithmic enforcement by online intermediaries with evidence from black box tinkering may provide invaluable insights into algorithmic copyright enforcement practices by online intermediaries. As the study demonstrates, black box tinkering allows us to check whether platforms consider fair use before automatically targeting questionable content.[158] This is especially interesting in relation to platforms that replace human judgment in detecting copyright infringement with automatic filtering algorithms.[159] Furthermore, black box tinkering may reveal important insights into the actual algorithmic implementation of fair use: How do algorithmic mechanisms of online

---

Schofield study had to take the responses of the surveyed platforms as given, without having any practical ability to actively challenge them.

157. Beyond surveying online platforms, Urban, Karaganis, and Schofield also provide a detailed quantitative examination of a random sample of takedown notices, taken from a set of over 108 million requests submitted to the Chilling Effects archive over a six-month period. Urban et al., *supra* note 32, at 80. Founded by the Berkman Center for Internet and Society at Harvard University, the Chilling Effects offers an invaluable clearinghouse for researchers and the public in general, making cease and desist letters concerning online content, and especially requests to remove content from online, publicly available. *Chilling Effects*, BERKMAN, https://cyber.harvard.edu/wg_home/chilling_effects (last visited Oct. 8, 2016). Urban, Karaganis, and Schofield's observational study provides very exciting findings about N&TD practices, including the domination of Google Web Search as the governing receiver of takedown requests; the domination of large entertainment companies as the ruling senders of takedown requests, and the domination of file-sharing sites as the key targets of most takedown requests; as well as their stimulating observations regarding the substantial and procedural validity of real-world notice-sending practices. *See* Urban et al., *supra* note 32, at 11, 70, 134–38. In his observational study, Seng analyzes half a million takedown notices and more than fifty million takedown requests. Seng, *supra* note 131. He examines the use and issuance of takedown notices by copyright owners and reporters and the response of service providers to them. *Id.* He further studies the relationship between the notices and requests and safe harbor provisions of the DMCA, and identifies ways in which the takedown process can be further improved to preserve the diversity and freedom of the internet. *Id.*

158. Again, observational studies are limited in their ability to examine ex ante implementations of online copyright enforcement because they are confined to analyzing the by-products of a given phenomenon—publicly available takedown notices, for instance. They do not have real-time access to unpublished, ex ante content restriction. *See supra* note 110 and accompanying text.

159. Such as YouTube's Content ID. *See* Ramon Lobato & Julian Thomas, *The Business of Anti-Piracy*, 6 INT'L J. COMMC'N 606, 612 (2012); *see also* Todd Spangler, *Vimeo Starts Scanning Videos for Copyright Violations*, VARIETY (May 21, 2014, 10:34 AM), https://variety.com/2014/digital/news/vimeostarts-scanning-videos-for-copyright-violations-1201188152/ (reporting on Vimeo's launch of Copyright Match). Note that the U.S. Court of Appeals for the 9th Circuit had recently made it clear that right-holders must consider fair use before sending a DMCA takedown notice, regardless of whether the notice is sent manually or automatically. *See* Lenz v. Universal Music Corp., 815 F.3d 1145, 1153 (9th Cir. 2016). This ruling may possibly suggest that automatic content-filtering technologies, which effectively stand at the position of right-holders in detecting copyright infringement, should also be required to consider fair use before targeting online content.

Electronic copy available at: https://ssrn.com/abstract=2741513

copyright enforcement effectively translate this four-factor discretional legal doctrine[160] into a workable copyright infringement detection algorithm? Further, black box tinkering can provide empirical evidence to support discrimination-related allegations against online intermediaries, claiming they afford their business partners preferable treatment, in contrast to small, independent creators.[161]

In short, black box tinkering may generate data that is cardinal in promoting public literacy in the black box systems that regulate online conduct. Furthermore, this data can improve the accountability of algorithmic copyright enforcement systems and generate social activism. From the perspective of online users, the ability to learn how automated filtering technologies determine copyright infringement may present them with better prediction capacities, which could ultimately enhance their notions of trust and fairness.[162] From the perspective of online platforms, knowing that users can effectively watch their practices and identify their possible misconduct may encourage platforms to improve their policies to better accommodate the interests of users.[163]

III. LEGAL CHALLENGES

By checking the hidden practices of enforcement algorithms, black box tinkering certainly offers important benefits in enhancing algorithmic accountability and promoting active public engagement, but it may still give rise to some legal challenges. This Part explains and attempts to rebut these challenges, relying on available legal doctrines. Lastly, this Part also proposes a possible measure of legal intervention to validate the use of black box tinkering and remove its potential legal barriers.

A. *Challenges Imposed by External Laws*

Applying black box tinkering methodology to study algorithmic enforcement may need to overcome various legal barriers. Specifically, tinkering with online systems of content adjudication may occasionally involve the submission of artificially generated content which deliberately—for the purpose of the research—violates applicable legal prohibitions, such as the prohibition against distribution of indecent

160. *See supra* note 155.

161. *See, e.g.*, Yafit Lev-Aretz, *Second Level Agreements*, 45 AKRON L. REV. 137, 184 (2012).

162. Perel & Elkin-Koren, *supra* note 12, at 519 (noting that because the general public does not understand law enforcement algorithms they are unable to hold algorithmic mechanisms of copyright enforcement accountable).

163. *Id.* at 525–26.

Electronic copy available at: https://ssrn.com/abstract=2741513

content to minors,[164] or the prohibition against dissemination of content containing copyright infringement.[165] Black box tinkering might even amount to violation of different "anti-hacking" and "computer intrusion" laws, such as the Computer Fraud and Abuse Act (CFAA),[166] the Federal Wiretap Act,[167] or the Stored Communications Act (SCA),[168] which may consider black box tinkering an unlawful intrusion into intermediaries' computer networks.[169]

For instance, under copyright law, uploading content that contains copyrighted material without prior authorization by the copyright owner may constitute copyright infringement.[170] Accordingly, insofar as black box tinkering studies involve uploading infringing material to online platforms to examine how they comply with copyright law (as in the black box tinkering study described in Part II), copyright law might discourage researchers from conducting this sort of black box tinkering and risking copyright liability.

Of course, researchers could escape copyright liability by obtaining the permission of copyright owners to use their protected material for experimental purposes, although some experiments may warrant the

164. Communications Decency Act, 47 U.S.C. § 230 (2012) (criminalizing the online provision of indecent materials to minors, unless the initiator had undertaken a good faith effort to determine the age of the person on the other end of the network).

165. Copyright Act of 1976, Pub. L. No. 94-553, 90 Stat. 2541 (codified as amended at 17 U.S.C. §§ 101–810 (2012)).

166. Pub. L. No. 98-473, 98 Stat. 2192 (1986) (codified as amended at 18 U.S.C. § 1030 (2012)).

167. Pub. L. No. 90-351, tit. III, 82 Stat. 197, 211–25 (1986) (codified as amended at 18 U.S.C. §§ 2510–22, 47 U.S.C. § 605).

168. 18 U.S.C. §§ 2701–12. The SCA was enacted as part of the Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848 (codified as amended in scattered sections of 18 U.S.C.)

169. *See* Perel & Elkin-Koren, *supra* note 12, at 523–24.

170. One of the sections of the Copyright Act of 1976 reads:

> Anyone who violates any of the exclusive rights of the copyright owner as provided by sections 106 through 122 or of the author as provided in section 106A(a), or who imports copies or phonorecords into the United States in violation of section 602, is an infringer of the copyright or right of the author, as the case may be.

17 U.S.C. § 501(a). Another section confers to the owner of copyright the exclusive rights to reproduce the copyrighted work, to prepare derivative works based upon the copyrighted work, to distribute copies of the copyrighted work, to perform the work publicly, and to display the work publicly. *Id.* § 106. Uploading content that contains copyrighted material may hence result in violation of the copyright owners' exclusive right to reproduce their work, to display it, to perform it, and to create derivative work from it.

Electronic copy available at: https://ssrn.com/abstract=2741513

copyright owner's ignorance.[171] Researchers may also rely on fair use,[172] which may apply to copying copyrighted materials for the experimental purpose of black box tinkering, provided it only made use of limited portions of the protected material as necessary for the research, in a transformative nonprofit manner, and with negligible potential harm to the market for the original work.[173]

However, as other copyright-related legal challenges might be harder to overcome, a research-related safe harbor is all the more necessary.[174] For instance, the DMCA codifies a right-verification procedure under which an eligible notification of claimed infringement must contain "a statement that the information in the notification is accurate, and under penalty of perjury, that the complaining party is authorized to act on behalf of the owner of an exclusive right that is allegedly infringed."[175] Many intermediaries that apply the DMCA N&TD procedure largely adhere to this statutory language and request that complainants sign similar declarations.[176] Yet investigating the hidden algorithmic implementations of N&TD by online intermediaries by uploading content containing copyrighted material, without prior authorization by relevant right-holders, may effectively require researchers to submit sham notices of copyright infringement on behalf of the true copyright owners.[177] Unfortunately, the DMCA sets no fair use-style exemption in this regard that may legitimize such conduct for pure research purposes and protect researchers from the risk of perjury.

B. *Challenges Imposed by Platforms' Terms-of-Use*

One of the greatest challenges in generating algorithmic oversight arises from its involving privately owned facilities and proprietary knowledge.[178] Consequently, applying black box tinkering methodology

171. For instance, if owners' authorization cannot be successfully obtained.

172. *See* 17 U.S.C. § 107.

173. Indeed, the black box tinkering study described uses extremely short portions of copyrighted material for experimentation purposes.

174. *See infra* Section III.C.

175. *See* 17 U.S.C. § 512(c)(3)(A).

176. See, for instance, Facebook's requirements for reporting an alleged infringement, *How Do I Report a Claim of Copyright Infringement?*, FACEBOOK, https://www.facebook.com/help/325058084212425?helpref=uf_permalink (last visited Oct. 9, 2016).

177. This is exactly what was done in the black box tinkering study described: After uploading the various types of content to online platforms, the researchers filed notices of takedown to remove the content, while recording the responses of the platforms. As emphasized earlier, because the study was conducted within Israel's jurisdiction, where the N&TD regime lacks the formalities of the U.S. DMCA N&TD, the researchers could not have been subject to this sort of copyright liability.

178. *See supra* notes 71–74 and accompanying text.

Electronic copy available at: https://ssrn.com/abstract=2741513

to study algorithmic enforcement by online intermediaries may often need to overcome different contractual barriers imposed by the examined platforms or software owners.[179] Particularly, any act of tinkering with platforms' online systems inevitably involves agreeing to its terms of use (ToU). While ToU are often considered contracts of adhesion that fail to satisfy the basic legal requirements of contract formation, addressing the legal validity of any such provisions in online intermediaries' ToU is beyond the scope of this Essay. Suffice it to say that provisions in ToU appear to be enforceable by courts,[180] and that such terms may set limits on what users—including researchers—are permitted to do when using online platforms.

For instance, in their ToU, intermediaries often explicitly prohibit users from using the platform for any unlawful, misleading, malicious, or discriminatory purpose.[181] However, black box tinkering may often reflect convergence of all these: most noticeably, challenging online systems of content adjudication requires researchers to submit superficial content which, although replicating real social media posts,[182] is nonetheless created specifically for the study's purpose. Because submitting such content arguably misleads both the platform and the users who consume the uploaded content, so that they believe the content is authentic, it may amount to misrepresentation.

Moreover, as explained earlier, in cases where tinkering involves violation of any legal provisions that apply to online content (e.g., laws against hate speech, discrimination, defamation, copyright infringement), it could amount to unlawful conduct that further violates the platforms' ToU.[183] For example, triggering online systems of algorithmic copyright

179. The tinkered technology may be subject to the TPMs or DRM that define the relationship between the owner of the copyrighted technology and its users. *See infra* notes 194–96 and accompanying text.

180. *See* Fteja v. Facebook, Inc., 841 F. Supp. 2d 829, 837, 841 (S.D.N.Y. 2012) (noting that "a number of courts . . . have enforced forum selection clauses in clickwrap agreements").

181. For instance, under Facebook's Statement of Rights and Responsibilities, users must make several safety obligations, including confirming not to do anything "unlawful, misleading, malicious, or discriminatory." *Statement of Rights and Responsibilities*, *supra* note 70. Similarly, Twitter's policies state that users "may not use our service for any unlawful purposes or in furtherance of illegal activities. International users agree to comply with all local laws regarding online conduct and acceptable content." *The Twitter Rules*, TWITTER, https://support.twitter.com/articles/18311# (last visited Oct. 9, 2016).

182. King et al., *supra* note 102, at 891.

183. For instance, Instagram states in its ToU that users "may not post violent, nude, partially nude, discriminatory, unlawful, infringing, hateful, pornographic or sexually suggestive photos or other content via the Service;" users are further prohibited from using "the Service for any illegal or unauthorized purpose," and they must "agree to comply with all laws, rules and regulations (for example, federal, state, local and provincial) applicable to [their] use of the Service and [their]

Electronic copy available at: https://ssrn.com/abstract=2741513

enforcement with infringing material—to the extent it is not fair use[184]—might violate those ToU that prohibit users from uploading content that violates the legitimate rights of third parties.[185]

In some cases the potential harm of applying black box tinkering methodologies to study the practices of algorithmic enforcement systems implemented by private entities would be negligible, ensuring that researchers engaging in such experimentation will not be exposed to any meaningful contractual liability. For instance, in the study of algorithmic N&TD practices described in Part II, the video snippets that were uploaded to test platforms' copyright enforcement practices were extremely short and did not contain any political or otherwise controversial content that could have harmed individuals' emotions, feelings, preferences, and the like. Hence, that specific application of black box tinkering research did not threaten any core human-related values, such as privacy and autonomy, but merely affected the content that online users could potentially consume.[186] Other black box tinkering studies, however, may be more offensive to third parties,[187] further justifying the need to set legal boundaries on experimental use within a safe harbor for black box tinkering.

Finally, black box tinkering may cause harm to the examined platform itself. Often, such harm would only be the negligible cost of adjudicating the artificial content submitted. Yet considering the robustness of algorithmic enforcement by online intermediaries,[188] a minimal increase in the number of content submissions that platforms confront is most likely economically insignificant. In other cases, however, the potential harm may be more substantial, further bolstering the need for legal intervention. This may happen, for instance, when a "denial of service" overload temporarily shuts down a site or when the content submitted is so harmful as to affect the site's reputation. The distribution of extremely

---

Content (defined below), including but not limited to, copyright laws." *Terms of Use,* INSTAGRAM, https://help.instagram.com/478745558852511 (last visited Oct. 9, 2016).

184. *See supra* notes 172–73 and accompanying text.

185. For instance, under Facebook's Statement of Rights and Responsibilities, users must commit not to "post content or take any action on Facebook that infringes or violates someone else's rights or otherwise violates the law." *Statement of Rights and Responsibilities*, *supra* note 70.

186. At least in relation to the study described, the researchers did not upload any political or otherwise controversial content they may have a deeper effect on humans' emotions.

187. Potential black box tinkering studies that test the practices of online content adjudication systems—especially in the contexts of child pornography and defamation—may impact core human values in a way that significantly exceeds the *de minimis* effect over users' content-consumption potential. For these particular cases, the development of a research safe harbor, of the type proposed below in Section III.C, may be critical.

188. *See supra* Section I.C.

Electronic copy available at: https://ssrn.com/abstract=2741513

harmful content submitted during a black box tinkering study may even give rise to actionable allegations against the platform for actively facilitating illegal conduct, such as incitement or defamation. Nevertheless, the identification of possible misconduct in the examined system's content-adjudication process resulting from black box tinkering experimentation, even with a direct economic impact on the platform's popularity, must be evaluated as an external social benefit of the research, not as intrinsic harm to the examined platform.

C. *Legal Intervention*

Because of the imperative need for black box tinkering to create a proper check on algorithmic enforcement by online intermediaries,[189] this Essay argues that the law should seek to encourage it. The legal barriers discussed above suggest that legal intervention might be necessary to ensure that researchers and social activists can apply and advance tinkering methodologies without risking legal liability. Generally, this Essay suggests that legislatures consider the enactment of black box tinkering safe harbors, designed to make researchers immune from liability for their intentional yet *de minimis* legal violations in the course of black box tinkering. Enacting statutory immunity would promote the active engagement of the public in revealing the hidden practices of the various algorithms that regulate humans' behavior, and encourage the public to review and eventually affect the way these algorithms function, while ultimately enhancing algorithmic accountability.

Specifically, in the context of copyright law this Essay proposes that the DMCA allow researchers to challenge how online platforms effectively implement the N&TD provisions, by guaranteeing a safe harbor for researchers who—when conducting their black box tinkering study—fail to comply with the technical formalities of the DMCA (especially the oath requirement),[190] or otherwise engage in copyright infringement.[191] This Essay further suggests that this exemption be contingent on researchers acting in good faith,[192] in terms both of diligently attempting to obtain prior authorization by right-holders to act on their behalf and of minimizing their study's potential harms.[193] This

189. *See supra* Section I.A.

190. *See supra* notes 172–73 and accompanying text.

191. *See supra* notes 170–71 and accompanying text.

192. Much like the DMCA exempts OSPs from liability for mistaken yet good faith removals of material. *See* 17 U.S.C. § 512(g)(1) (2012).

193. For instance, to minimize the possible legal implications arising from platforms' potential deception and misrepresentation caused by black box tinkering research, researchers may consult the varied ethical guidelines that are available to help researchers decide whether deception is justified in their research. *See, e.g.*, Herbert C. Kelman, *Human Use of Human Subjects: The Problem of Deception*

Electronic copy available at: https://ssrn.com/abstract=2741513

Essay also proposes that researchers obtain prior approval by the relevant ethical review boards to conduct their research.

Note that reverse-engineering exemptions are not new to copyright law, which had previously recognized the need to grant some leeway for tinkering with specific regulating technologies that enforce Digital Rights Management (DRMs) and Technology Protection Measures (TPMs). These technologies essentially reflect different types of encrypted computer codes that are incorporated into fixations of copyrighted material, such as DVDs or music files, to prevent illegal copying and public distribution of copyrighted works.[194] Because DRMs and TPMs—like any other technology—are not tamper-proof, defenders of strong copyright called for prohibition of the use, development, or distribution of technologies designed to "circumvent" (e.g., hack, crack, or break) such access control technologies.[195] In response to these pressures, the legislature mandated the DMCA's anti-circumvention provisions,[196]

---

in Social Psychological Experiments, 67 PSYCHOL. BULL. 1, 8 (1967) (suggesting ways that social psychologists can deal with concerns about use of deception in experiments); Alan C. Elms, *Keeping Deception Honest: Justifying Conditions for Social Scientific Research Stratagems*, ALAN C. ELMS VIRTUAL LIBR., http://www.ulmus.net/ace/library/keepingdeceptionhonest.html (last visited Oct. 9, 2016). One popular set of guidelines is provided by Pascal, Leone, Singh, and Scoboria, who suggest that researchers mainly: (1) assure that all reasonably possible costs and benefits have been accounted for, (2) ensure the study cannot be done either without or with a lesser degree of deception, (3) find out whether the deception is associated with more than minimal risk, and (4) affirm that no possible risks have been overlooked in the description of the study. *See* ROGER D. WIMMER & JOSEPH R. DOMINICK, MASS MEDIA RESEARCH 73–74 (10th ed. 2014). Furthermore, after conducting a black box tinkering study that involves the minimal degree of deception required for achieving the study's objectives, it is important to debrief the examined platforms about the study. That is, researchers should thoroughly describe the purpose of the study, explain the use of deception, and encourage the study's subjects to ask questions about the study. *See id.* at 74.

194. Stephen M. Kramarsky, *Copyright Enforcement in the Internet Age: The Law and Technology of Digital Rights Management*, 11 DEPAUL-LCA J. ART & ENT. L. 1, 10–13 (2001).

195. *See* Samuelson, *supra* note 118, at 521, 547.

196. 17 U.S.C. § 1201. The most pertinent of the DMCA's anti-circumvention provisions read in part:

> (a) Violations Regarding Circumvention of Technological Measures.—
>
> (1)(A) No person shall circumvent a technological measure that effectively controls access to a work protected under this title.
>
> . . .
>
> (2) No person shall manufacture, import, offer to the public, provide, or otherwise traffic in any technology, product, service, device, component, or part thereof, that—

Electronic copy available at: https://ssrn.com/abstract=2741513

which impose serious limitations on the act of tinkering with DRMs and TPMs.

Nevertheless, alongside the anti-circumvention restricting provisions came specific narrow exceptions, which maintain some extent of freedom to tinker, especially in relation to achieving program-to-program interoperability or engaging in encryption research and computer security testing.[197] Accordingly, if the importance of specific acts of tinkering was acknowledged for technologies that restrict access to distinct content, all the more should it be recognized for technologies that control prominent content distribution channels and effectively shape public discourse.[198] As we explained elsewhere:

> Having the ability to circumvent mechanisms of algorithmic copyright enforcement thus reaches beyond the narrow interests of lawful owners of specific copies of copyrighted content and curious technologists because it enables users to

---

> (A) is primarily designed or produced for the purpose of circumventing . . . ;
>
> (B) has only limited commercially significant purpose or use other than to circumvent . . . ; or
>
> (C) is marketed by that person or another acting in concert with that person with that person's knowledge for use in circumventing . . . .
>
> . . .
>
> (b) Additional Violations.—
>
> (1) No person shall manufacture, import, offer to the public, provide, or otherwise traffic in any technology, product, service, device, component, or part thereof, that—
>
> (A) is primarily designed or produced for the purpose of circumventing protection . . . ;
>
> (B) has only limited commercially significant purpose or use other than to circumvent protection . . . ; or
>
> (C) is marketed by that person or another acting in concert with that person with that person's knowledge for use in circumventing protection . . . .

Id.

197. For instance, 17 U.S.C. § 1201(g)(1)(A) authorizes "encryption research"—"activities necessary to identify and analyze flaws and vulnerabilities of encryption technologies applied to copyrighted works, if these activities are conducted to advance the state of knowledge in the field of encryption technology or to assist in the development of encryption products."

198. *See* Michael S. Sawyer, *Filters, Fair Use & Feedback: User-Generated Content Principles and the DMCA*, 24 BERKELEY TECH. L.J. 363, 380–82 (2009); *see also* Perel & Elkin-Koren, *supra* note 12, at 521–22.

Electronic copy available at: https://ssrn.com/abstract=2741513

> contest what some scholars have characterized as the "privication" of information that would have otherwise been public.[199]

Furthermore, advancing such a reverse-engineering safe harbor also accords with other IP-related legal doctrines which—like the exemptions to the anti-circumvention provisions under the DMCA—legitimize specific acts of tinkering with existing artifacts. For instance, trade secrecy law regards reverse engineering as a lawful way to acquire knowhow that the product's manufacturer may claim as a trade secret,[200] allowing "people or firms to buy products, disassemble them, study their components, and test them in various ways to figure out how they work."[201] Another important doctrine in IP law that fosters freedom to tinker is known as the "first sale" limit on IP rights.[202] Essentially, this doctrine "allows those who have acquired products in the marketplace considerable freedom to use, modify, and resell [them] as they wish, even if the products are protected in whole or in part by IP rights."[203] Patent law further facilitates freedom to tinker under the experimental use exception, which permits the use of another's patented device, when such use is for philosophical inquiry, curiosity, or amusement.[204] Hence, research institutions can generally use patented devices without authorization for non-commercial, experimental purposes, made outside the ordinary course of business.

Indeed, freedom to tinker is a settled concept. It is a well-known expression of humans' natural curiosity, inquisitiveness, and independent

---

199. Perel & Elkin-Koren, *supra* note 12, at 521–22 & n.300 ("Privication describes the possibility of private publication, where content providers distribute content on a large-scale but at the same time retain control over access."); *see also* Jonathan Zittrain, *What the Publisher Can Teach the Patient: Intellectual Property and Privacy in an Era of Trusted Privication*, 52 STAN. L. REV. 1201, 1218 (2000).

200. *See, e.g.*, Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 476, 490 (1974).

201. Samuelson, *supra* note 98, at 5.

202. *Id.* at 6.

203. *Id.*

204. *See, e.g.*, Embrex, Inc. v. Serv. Eng'g Corp., 216 F.3d 1343, 1349 (Fed. Cir. 2000). While the application of this exemption to academic research institutions was challenged in *Madey v. Duke University*, 307 F.3d 1351, 1353, 1362 (Fed. Cir. 2002)—mainly because in that case the use of the patent by the university clearly furthered its legitimate business interests, including educating and enlightening students and faculty—the vast majority of scholars have criticized that ruling. *See, e.g.*, Natalie M. Derzko, *In Search of a Compromised Solution to the Problem Arising from Patenting Biomedical Research Tools*, 20 SANTA CLARA COMPUTER & HIGH TECH. L.J. 347, 365–66 (2004); Katherine J. Strandburg, *What Does the Public Get? Experimental Use and the Patent Bargain*, 2004 WIS. L. REV. 81, 84–85; Andrew J. Caruso, Comment, *The Experimental Use Exception: An Experimentalist's View*, 14 ALB. L.J. SCI. & TECH. 215, 220 (2003); Kevin Sandstrom, Note, *How Much Do We Value Research and Development?: Broadening the Experimental Use Exemption to Patent Infringement in Light of* Integra Lifesciences I, Ltd. v. Merck KGaA, 30 WM. MITCHELL L. REV. 1059, 1067 (2004).

Electronic copy available at: https://ssrn.com/abstract=2741513

thought. When non-transparent code replaces human judgment, the ability to tinker with the code serves additional social objectives. Beyond providing individuals with practical ability to interact with the hidden algorithms that regulate their behavior, freedom to tinker enables a real check on the practices of influential black boxes. It encourages social activism and exposes algorithms to real and effective criticism that further enhances their trustworthiness. Accordingly, policy makers should promote the use of black box tinkering and make every effort to remove the barriers that diminish its approachability.

CONCLUSION

Presently, as hidden algorithms get to control more and more aspects of everyday conduct, including managing online behavior and enforcing legal rights, it is crucial to subject them to adequate scrutiny. As passive, transparency-driven observations of algorithmic enforcement systems are limited in their capacity to check the practices of non-transparent, constantly evolving algorithms, it is essential to encourage the active engagement of the public in challenging unknown and possibly biased systems of algorithmic governance. A proactive methodology of black box tinkering enables researchers to challenge the black box systems around them and reveal their misconduct. While this Essay has paid special attention to algorithmic enforcement of copyright law, its general insights are entirely relevant to other manifestations of algorithmic enforcement—of criminal law or civil law, by private entities or by governmental agencies. First, all enforcing algorithms share the same inherent characteristics of non-transparency and machine learning, which curtail their ability to be sufficiently predictable and discernible. Second, the automatic nature of algorithmic enforcement systems makes them ubiquitous, reinforcing the problem of inspecting voluminous amounts of transparent data. Third, challenging algorithmic enforcement systems through black box tinkering inevitably involves some degree of deception or misrepresentation, which may raise legal challenges. Accordingly, this Essay concludes that black box tinkering is a valuable research methodology that deserves more attention in social science studies that explore the practices of hidden technologies. As a first step, this Essay therefore calls for the removal of all legal barriers that may discourage researchers from exploiting this methodology and benefitting from its valuable advantages.

Electronic copy available at: https://ssrn.com/abstract=2741513

Electronic copy available at: https://ssrn.com/abstract=2741513