LAW IN, LAW OUT: LEGALISTIC FILTER BUBBLES AND THE ALGORITHMIC PREVENTION OF NONCONSENSUAL PORNOGRAPHY

Daniel Maggen[†]

In 2019, Facebook announced that it had begun using machine-learning algorithms to preemptively screen uploads for nonconsensual pornography. Although the use of screening algorithms has become commonplace, this seemingly minor move from reactive to preemptive legal analysis–based prevention—this Article argues—is part of a groundbreaking shift in the meaning and effect of algorithmic screening, with potentially far-reaching implications for legal discourse and development.

To flush out the meaning of this shift, the Article draws on the filter bubble theory. Thus far, the phenomenon of filter bubbles has been synonymous with personalized filtering and the social polarization and radicalization it is prone to producing. Generalizing on this idea, the Article suggests that the control filtering algorithms have over the information brought before users can shape users' worldviews in accordance with the algorithm's measure of relevance. Algorithmic filtering produces this effect by enhancing users' trust in the applicability of the measure of relevance and "invisibly hiding" any information that conflicts with it. The Article argues that, in the case of filtering algorithms that use a legal classification as their measure of relevance, the result is a legalistic filter bubble that can essentialize dominant legal paradigms and suppress information that challenges

[†] Visiting Fellow, Information Society Project at Yale Law School; Lecturer, Yale University. This Article benefited immensely in the various stages of its development from conversations with Gilad Abiri, Jack Balkin, Kiel Brennan-Marquez, William Eskridge, Joshua Fairfield, Paul Kahn, Asaf Lubin, Daniel Markovitz, Frank Pasquale, Lawrence Solum, Alicia Solow-Niederman, and Christina Spiesel, as well as from participants of the Information Society Project workshop, the Law & Technology Virtual Workshop at The Nebraska Governance & Technology Center, and the Eighth Tel Aviv Privacy, Cyber, and Technology Workshop at the Edmond J. Safra Center for Ethics at Tel Aviv University. I am also indebted to the students in my "Crime in the 21st Century" seminar at Yale University for their thought-provoking discussions of some of this Article's themes. Finally, I want to thank the editors at *Cardozo Law Review* for their many thoughtful comments and suggestions.

their usefulness and decency. These effects, the Article suggests, can significantly impede legal evolution as they drive a wedge between adjudication and the greater normative universe it inhabits.

In the case of filtering algorithms that use the legal category of nonconsent as their measure of relevance, the emergence of a filter bubble will effectively cement nonconsent as the gravamen of violative distribution and insulate decision-makers from exposure to consensual harms. Although the Article does not suggest that we ban consensual but harmful distribution of sexual materials, it argues that the emergence of a filter bubble can hinder the development of a vibrant normative debate on the meaning of sexual autonomy.

Table of Contents

INTRODUCTION

A man uses a cell phone to record a video of a woman and later uploads the video to a private Facebook group.[1] When the upload is complete, a popup screen notifies the man that a machine learning–powered algorithm detected sexually explicit images in the video and that further algorithmic analysis determined that the upload appears to be without the woman's consent.[2] The popup informs the man that a final decision is awaiting human review and that he may provide evidence to establish consent. It also notifies him that if final review determines that the upload was nonconsensual, Facebook will suspend his account, make efforts to prevent any distribution of the video, and attempt to notify the victim as well as the authorities.[3]

This narrative is based on Facebook's 2019 announcement that it had begun using machine-learning algorithms to automatically screen uploads for nonconsensual distribution of intimate images.[4] Although this announcement drew little attention, it is no less than groundbreaking and one of the most significant steps thus far in the rise of *legalistic* filtering, meaning the use of algorithms to independently determine a piece of information's legal classification.[5] Until recently,

[1] The choice to use gendered language reflects the highly gendered nature of nonconsensual pornography. For empirical data, see ASIA A. EATON, HOLLY JACOBS & YANET RUVALCABA, 2017 NATIONWIDE ONLINE STUDY OF NONCONSENSUAL PORN VICTIMIZATION AND PERPETRATION: A SUMMARY REPORT 12 (2017), https://www.cybercivilrights.org/wp-content/uploads/2017/06/CCRI-2017-Research-Report.pdf [https://perma.cc/L2CT-UVM4] (finding that women are 1.7 times likelier to be victims than men); AMANDA LENHART, MICHELE YBARRA & MYESHIA PRICE-FEENEY, NONCONSENSUAL IMAGE SHARING: ONE IN 25 AMERICANS HAS BEEN A VICTIM OF "REVENGE PORN" 5 (2016), https://datasociety.net/wp-content/uploads/2016/12/Nonconsensual_Image_Sharing_2016.pdf [https://perma.cc/Q4GZ-HAHU] (finding that the phenomenon is much more prevalent for female victims); Abby Whitmarsh, *Analysis of 28 Days of Data Scraped from a Revenge Pornography Website.*, WORDPRESS.COM (Apr. 13, 2015), https://everlastingstudent.wordpress.com/2015/04/13/analysis-of-28-days-of-data-scraped-from-a-revenge-pornography-website [https://perma.cc/2KF2-HCJ3] (finding revenge porn to overwhelmingly target women).

[2] This scenario is inspired by information provided by Facebook on this procedure. *See* Antigone Davis, *Detecting Non-Consensual Intimate Images and Supporting Victims*, META (Mar. 15, 2019), https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images [https://perma.cc/HCE3-4PVW]. More details are provided *infra* Section III.C.

[3] The notification component is not part of Facebook's announced policy. It can, however, accord with a more general duty to report certain crimes. *Cf.* Alexander Tsesis, *Social Media Accountability for Terrorist Propaganda*, 86 FORDHAM L. REV. 605 (2017) (discussing platforms' duty to report terrorism-related crimes and child abuse).

[4] *See* Davis, *supra* note 2.

[5] *See* Davis, *supra* note 2; *see also* Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 LEWIS & CLARK L. REV. 857, 885 (2020) (describing the turn to algorithmic content moderation as a "sea change . . . . transforming the way laws govern the public sphere").

algorithms have helped human decision-makers make legal decisions in various ways, including through risk assessment, fact-finding, and other forms of evaluation auxiliary to legal analysis.[6] However, recent years have seen the rise of algorithms that emulate legal analysis to determine whether information is worthy of human decision-makers' attention.[7] The case of nonconsensuality detection is not entirely unique in this emerging trend, but it stands out for its adherence to a distinct legal category, the independence of its legal analysis, and the extent of its control over which information is brought before human decision-makers. While in past use cases an algorithm's operation relied on user and other input, in the case of nonconsensual-pornography filtering and other systems like it, the algorithm *itself* is tasked with modeling the meaning of the legal category it screens and implementing this model on previously unscrutinized information. This Article is the first to discuss this emerging trend and the effects it is prone to have on legal development, as well as the first to note its potential influence on the

[6] On such decision-assisting systems, see Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 663 (2017); Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 2021–22 (2017); RASHIDA RICHARDSON, JASON M. SCHULTZ & VINCENT M. SOUTHERLAND, AI NOW INST., LITIGATING ALGORITHMS 2019 US REPORT: NEW CHALLENGES TO GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS (2019); Elizabeth E. Joh, *The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing*, 10 HARV. L. & POL'Y REV. 15 (2016); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871 (2016); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 101 (2014) [hereinafter Surden, *Machine Learning*] (discussing algorithms in the service of lawyers); Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305 (2019) [hereinafter Surden, *Artificial Intelligence*] (discussing algorithm making legal predictions).

[7] Similar legalistic algorithmic screening is increasingly used to cut down in size cases brought before human decision-makers in the context of copyright infringements, toxic content on social media, child maltreatment, insider trading, policing, regulatory compliance, and the list grows in length with every passing day. *See* Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473 (2016) (discussing copyright infringement); Niva Elkin-Koren, *Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence*, BIG DATA & SOC'Y, July–Dec. 2020 (discussing toxic content on social media); Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001 (2019) (discussing toxic content on social media); Daniel Maggen, *Predict and Suspect: The Emergence of Artificial Legal Meaning*, 23 N.C. J.L. & TECH. 67 (2021) (discussing child maltreatment); Todd Ehret, *SEC's Advanced Data Analytics Helps Detect Even the Smallest Illicit Market Activity*, REUTERS (June 30, 2017, 1:11 PM), https://www.reuters.com/article/bc-finreg-data-analytics/secs-advanced-data-analytics-helps-detect-even-the-smallest-illicit-market-activity-idUSKBN19L28C [https://perma.cc/C5Y3-YUDX] (discussing insider trading); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327 (2015) (discussing policing); Joh, *supra* note 6 (discussing policing); Rich, *supra* note 6 (discussing policing); Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669 (2010) (discussing regulatory compliance). For a general overview, see DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020).

way we come to think of the harms of nonconsensual and unwelcome distribution of sexual materials.[8]

To understand the significance of the shift to legalistic filtering, we must turn to the familiar discussion of the emergence of so-called algorithmic "filter bubbles."[9] This term has become synonymous with the harms of personalized algorithmic filtering and the polarization and radicalization that can result from personalized filter bubbles.[10] However, if we take a step back from this fixation on personalized filtering, we can understand filter bubbles as involving three generally applicable ideas. First, the basic premise of the filter bubble theory is that the use of algorithms can have profound adverse effects even when the algorithm is doing precisely what it is supposed to do. Hence, although the lion's share of legal scholarship criticizing the use of algorithms focuses on their potential erroneousness and biases,[11] centering on these failures risks obfuscating issues inherent in filtering algorithms as such.

The second insight of the filter bubble theory is that the algorithm's control over the flow of information can cause a considerable winnowing effect, constricting decision-makers' worldviews.[12] The

[8] Similar concerns have been raised by Niva Elkin-Koren, who focuses on the effect of filtering on speech regulation and democratic deliberation. *See* Elkin-Koren, *supra* note 7. Elkin-Koren suggests responding to these challenges by introducing adversarial algorithmic systems, a response that could potentially assist in dealing with legalistic filter bubbles as well.

[9] *See infra* Section II.A. For a general discussion of the filter bubble theory, see ELI PARISER, THE FILTER BUBBLE: WHAT THE INTERNET IS HIDING FROM YOU (2011).

[10] *See, e.g.*, Engin Bozdag, *Bias in Algorithmic Filtering and Personalization*, 15 ETHICS & INFO. TECH. 209 (2013) (discussing personalized filter bubbles); Tarleton Gillespie, *The Relevance of Algorithms*, *in* MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY 167, 167 (Tarleton Gillespie, Pablo J. Boczkowski & Kirsten A. Foot eds., 2014) (discussing personalized search engines and other personalized filters); PARISER, *supra* note 9, at 15 (focusing on personalized filtering). I am indebted to Elana Zeide for pointing out the connection to the filter bubble scholarship.

[11] *See, e.g.*, Bamberger, *supra* note 7, at 676 (discussing how the use of algorithms can distort regulatory compliance); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 415 (2017) (emphasizing the salience of risk of algorithmic inaccuracies); Evelyn Douek, *Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 774–75 (2021) (discussing the erroneous implementation of Facebook's ban on nudity). On hidden biases, see Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Rashida Richardson, Jason M. Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15 (2019).

[12] *See* Bamberger, *supra* note 7, at 676 ("[T]echnology systems are not merely tools for implementing the goals of those who employ them; they shape the meaning of those goals themselves. In Heidegger's words, they create a Gestell, or world view, that alters the perceptions of the decisionmakers they inform."); Gillespie, *supra* note 10, at 167, 187 ("Algorithms play an

shape this constriction takes derives from the measure used by the filtering algorithm to determine the relevance of any piece of information.[13] The (personalized) filter bubble theory thus draws attention to a decision made by Google in 2009 to change its search engine's measure of relevance such that instead of measuring a web page's *user-neutral* relevance to the search terms, it began to measure a page's relevance to the *specific user's* preferences, as inferred from information obtained by Google.[14] As the filter bubble theory suggests, this seemingly innocuous shift in how the algorithm measures relevance revolutionized the information-consumption habits of all of the search engine's users, ensnaring them in personalized echo chambers.[15]

Once we generalize from personalized filtering, the importance of the shift to a legalistic measure of relevance becomes self-evident. Like the shift to personalized filtering at the time, the contemporary turn to legal classifications as the measures of relevance in normative matters ushers in a new age of legalistic filter bubbles. Like its personalized counterpart, legalistic filtering is prone to constricting the worldviews of those reliant on it and making them congruent with the legal classifications informing the algorithm's design.

The third insight of the filter bubble theory is that filter bubbles have the dual effects of entrenching users' preexisting tendencies to accept the relevance of the filtering criteria and "invisibly hiding" information that contradicts them.[16] In the case of personalized filtering, this involves fostering users' confirmation biases and preventing encounters with conflicting preferences, which become

increasingly important role in selecting what information is considered most relevant to us . . . ."); PARISER, *supra* note 9, at 21–46, 82 ("Like a lens, the filter bubble invisibly transforms the world we experience by controlling what we see and don't see.").

[13] *See* Gillespie, *supra* note 10, at 179 ("Each algorithm is premised on both an assumption about the proper assessment of relevance, and an instantiation of that assumption into a technique for (computational) evaluation.").

[14] *See* PARISER, *supra* note 9, at 1–3 (discussing Google's announcement); Bryan Horling & Matthew Kulick, *Personalized Search for Everyone*, GOOGLE: OFFICIAL BLOG (Dec. 4, 2009), https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html [https://perma.cc/GFB2-2EGW].

[15] On the radicalization this can produce, see CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 179–97 (2016); Zeynep Tufekci, *YouTube, the Great Radicalizer*, N.Y. TIMES (Mar. 10, 2018), https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html [https://perma.cc/F4UB-A449].

[16] *See* PARISER, *supra* note 9, at 91 ("Because the filter bubble hides things invisibly, we're not as compelled to learn about what we don't know.").

"unknown unknowns."[17] Likewise, in legalistic filtering, using a legal category as the measure of relevance entails reaffirming decision-makers' trust in the applicability of the legal paradigms informing this category and invisibly hiding information that undermines these paradigms.[18]

Undoubtedly, path dependence has always been a feature of legal adjudication.[19] Once a legal classification sets in, it can be difficult for lawyers to question it. However, healthy legal systems are at least occasionally capable of reevaluating the appropriateness and decency of their dominant paradigms.[20] Such reevaluation and reflection are often occasioned by the accumulation of a critical mass of encounters that puts the wisdom or legitimacy of dominant norms into question.[21] Legalistic filter bubbles, however, can make these occasions for reflection few and far between, preventing the emergence of a critical mass and making debate and reflection much less likely.

This danger becomes apparent as legalistic algorithms are enlisted in the fight against nonconsensual pornography. Nonconsensual pornography, previously referred to as "revenge porn" and today also described more appropriately as image-based sexual abuse, involves the distribution of sexual materials without the consent of the persons appearing in them.[22] Using algorithms to model and apply the legal meaning of nonconsent is prone to creating a filter bubble that could

[17] *See* Gillespie, *supra* note 10, at 187 ("Google's solution is operationalized into a tool that billions of people use every day, most of whom experience it as something that simply, and unproblematically, 'works.'"); PARISER, *supra* note 9, at 88–89, 106 ("In the filter bubble . . . . [y]ou don't see the things that don't interest you at all. You're not even latently aware that there are major events and ideas you're missing.").

[18] *See* Gillespie, *supra* note 10, at 172 ("The particular patterns whereby information is either excluded from a database, or included and then managed in particular ways . . . . help establish and confirm standards of viable debate, legitimacy, and decorum."); PARISER, *supra* note 9, at 84 ("First, the filter bubble surrounds us with ideas with which we're already familiar (and already agree), making us overconfident in our mental frameworks. Second, it removes from our environment some of the key prompts that make us want to learn.").

[19] *See generally* Oona A. Hathaway, *Path Dependence in the Law: The Course and Pattern of Legal Change in a Common Law System*, 86 IOWA L. REV. 601 (2001).

[20] To use Joshua Fairfield's example, at some point, law is forced to recognize that the right way to adjudicate slavery is to dispose of this legal category altogether. JOSHUA A.T. FAIRFIELD, RUNAWAY TECHNOLOGY: CAN LAW KEEP UP? 63 (2021); *see also id.* at 67–68, 79–81, 136 (discussing the need for law to adapt to social changes). This idea has been most forcefully argued in Robert M. Cover, *Foreword:* Nomos *and Narrative*, 97 HARV. L. REV. 4 (1983).

[21] *See* Hathaway, *supra* note 19, at 641–42 (juxtaposing legal and biological evolution).

[22] *See, e.g.*, *What to Do if You're the Target of Revenge Porn*, FED. TRADE COMM'N CONSUMER ADVICE, https://www.consumer.ftc.gov/articles/what-do-if-youre-target-revenge-porn [https://perma.cc/B2QB-7JQ5] (discussing "revenge porn"); Clare McGlynn & Erika Rackley, *Image-Based Sexual Abuse*, 37 OXFORD J. LEGAL STUD. 534 (2017) (discussing image-based sexual abuse).

suppress any debate on whether the consent paradigm is the appropriate way of protecting sexual autonomy. Although there can be no dispute that nonconsensual distribution of sexual materials must be prohibited, formal nonconsensuality does not exhaust the destructive effect that unwelcome but formally consensual distribution can have on the victim's sexual autonomy and well-being.[23] A legalistic filter bubble can obscure this simple fact by essentializing nonconsent as the gravamen of violative distribution and invisibly hiding the harms of consensual distribution. Although this Article does not directly take sides in the debate on the appropriateness of the consent paradigm,[24] it argues that such debates are vital to the vitality of our normative and legal environments.[25]

The Article proceeds in three distinct Parts. Part I sets the context for the discussion by distinguishing the challenge posed by legalistic filter bubbles from familiar criticisms. It does so by exploring some of the possible reasons for legal scholarship having thus far overlooked the rise of legalistic filtering. Part II presents and elaborates on the filter bubble theory, using it to explain the vital importance of the shift to legalistic filtering and its potential effects. Part III then outlines the potential emergence of a legalistic filter bubble in the prevention of nonconsensual pornography. This bubble, this Part argues, can entrench a transactional understanding of sexual autonomy and obscure the existence of consensual harms.

I. OVERLOOKING THE TURN TO LEGALISTIC FILTERING

Critical attention to the legal implications of decision-making algorithms has thus far been mostly split between the risks posed by faulty algorithms and the harms of personalized filtering.[26] The harms

[23] As the 2015 documentary *Hot Girls Wanted*, coproduced by Mary Anne Franks, vividly demonstrates, there is often very little daylight between nonconsensual and formally consensual pornographic images uploaded online; both can equally disregard the sexual autonomy of those depicted, both can be equally harmful to their well-being. *See* HOT GIRLS WANTED (Two to Tangle Productions 2015).

[24] Elsewhere I suggested that we should consider sexual degradation as an alternative to the consent paradigm. *See* Daniel Maggen, *"When You're a Star": The Unnamed Wrong of Sexual Degradation*, 109 GEO. L.J. 581 (2021).

[25] *See infra* Part III.

[26] In addition, an important trend academia and industry have been exploring is the fairness, accountability, transparency, and ethics (FATE) considerations that can affect the legitimacy of algorithmic decision-making, closer in its trajectory to the filter bubble approach. For discussion of these considerations, see Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103 (2018); Sarah Valentine, *Impoverished Algorithms:*

of legalistic filtering involve neither of the two, which perhaps explains why it has received so little attention.[27] Accordingly, to better understand the nature of this neglected challenge, this Part will set the stage for the following discussion by addressing some of the dominant criticisms leveled at algorithmic decision-making, distinguishing them from the challenge posed by legalistic filtering. After a brief outline of machine learning, the technology driving many of these algorithms and many of their failings, this Part will ask why, if legalistic filter bubbles indeed pose such a considerable challenge to legal evolution, the turn to legalistic filtering has received little to no attention.

I will suggest three possible reasons for this oversight. The first concerns the framing of the question. Machine-learning technology is known to introduce hidden biases into the algorithm's operation. As we shall see, focusing on the design features that produce such problems can obscure the significance of the algorithm's measure of relevance and its role in the algorithm's creation, making it appear to be an innocuous aspect of the algorithm's design.

The second reason concerns the sense of urgency produced by algorithmic errors, coupled with the exigency of some of the purposes for which the algorithms are used. As illustrated by the fight against nonconsensual pornography, algorithms can be an invaluable and even indispensable instrument in the prevention of grave harms. In such cases, the primary concern is ensuring that the algorithm does what it is supposed to; structural issues like the filter bubble, which are only fully manifested in the longer run, are overshadowed.

The third reason, and perhaps the most salient one, is the notion that in legal or law-adjacent decision-making, there can be little wrong with an instrument that correctly implements the applicable norm. Although many authors have noted the nuanced ways in which algorithms can fail to live up to this grandiose expectation, they seem to share an implicit assumption that if algorithms were capable of meeting it, they would be beyond reproach.

A. *The Ins and Outs of Machine Learning*

Generally speaking, using machine-learning algorithms involves two distinct stages: creating the model animating the algorithm and

Misguided Governments, Flawed Technologies, and Social Control, 46 FORDHAM URB. L.J. 364 (2019).

[27] Notable exceptions are Elkin-Koren, *supra* note 7, at 3–5; Tarleton Gillespie, *Content Moderation, AI, and the Question of Scale*, BIG DATA & SOC'Y, July–Dec. 2020, at 3–4 (briefly discussing this challenge in general terms).

running the algorithm.[28] Until recently, creating an algorithmic model involved manually representing the task the algorithm was to perform in formal, logic-based instructions, and coding them into computer-comprehendible operations.[29] Manually creating a translation algorithm, for instance, required developing formal models of meaning for each word or phrase in each language and creating the computer instructions that matched the corresponding sets of meanings.[30] However, since natural language is often context dependent, a word's meaning can be resistant to formal representation, making manual modeling extremely laborious.[31] Machine-learning technology revolutionized this process; instead of manually creating language models, Google's learning algorithm, for instance, discerns the connection between texts in different languages by *automatically* inferring it from the vast volume of digitized books published in different languages in Google's possession.[32] Although machine-learning algorithms vary in their purposes and methodologies, they generally follow the same move from manual to automatic modeling, although, in practice, any single algorithm's design is rarely reliant solely on automated learning.[33]

The learning algorithm automatically infers the rules composing the running model through an iterative process commonly referred to as "training."[34] In training, the learning algorithm models a database of examples to attain the hidden principles that governed whatever operation created the data.[35] Hence, a machine-learning algorithm used to determine whether the uploading of a sexual video is consensual will involve training on past decisions in an attempt to attain from the

[28] *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 655 (2017) (discussing the two stages of machine learning).

[29] *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 22–24 (4th ed. 2021) (discussing non-machine-learning artificial intelligence).

[30] *See* JOHN D. KELLEHER, DEEP LEARNING 22–30 (2019) (discussing the key ingredients of machine learning).

[31] *See* NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE: A HISTORY OF IDEAS AND ACHIEVEMENTS 354–61 (2009) (discussing the history of semantic networks).

[32] *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 50–72 (2013) (discussing the rise of machine learning–based modeling).

[33] For a general discussion of the role of data scientists in machine learning, see JOHN D. KELLEHER & BRENDAN TIERNEY, DATA SCIENCE 35–36 (2018).

[34] *See* RUSSELL & NORVIG, *supra* note 29, at 693 (discussing learning in general).

[35] *See* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 97 (2016) (discussing training through optimization).

training data the principles that informed these decisions.[36] To be used in the training process, information in these datasets will be separated into "input data" and "output data," with input datapoints including the different features of each example and the output datapoints composing the decision made in each case.[37] In *supervised* machine learning, which is the most common method of producing algorithms capable of legal classification, these decisions, *qua* output variables, are the "labels" "supervising" the training process.[38]

The training process aims to extract from patterns in the data the mathematical function that presumably connects the input and output data.[39] Extracting the function is performed by iteratively altering the relationship between the input and output variables in the algorithmic model, randomly or according to predetermined heuristics, and measuring "fitness," meaning the distance between the model and the desired function it seeks to infer from the data.[40] In supervised learning, the evolving model iteratively changes in training by selecting, arranging, and assigning different weights to input variables, while the output variables remain relatively constant, anchoring the training process as proxies for the ground truth it seeks to emulate.[41] Measuring the fitness of each iteration and using it to extract from the data the function that connects input and output variables is the secret sauce of machine learning, the key to its ability to emulate naturally occurring phenomena and human behavior. Although considerable designer intervention is inevitable, it is the automatic attainment of the function inherent in the data that drives machine-learning technology.[42]

[36] *See* Olivia Solon, *Inside Facebook's Efforts to Stop Revenge Porn Before It Spreads*, NBC NEWS (Nov. 19, 2019, 11:15 AM), https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631 [https://perma.cc/5LW9-RDGY].

[37] *See* Lehr & Ohm, *supra* note 28, at 665–66 (discussing input and output variables).

[38] *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 35, at 102–04 (discussing how labels supervise learning).

[39] On this crucial process, see KELLEHER, *supra* note 30, at 6–12, 26; Lehr & Ohm, *supra* note 28, at 677.

[40] The idea of "fitness" is specifically used in "evolutionary" methods of learning. *See* Y. Jin, *A Comprehensive Survey of Fitness Approximation in Evolutionary Computation*, 9 SOFT COMPUTING 3 (2005). More generally it is referred to as the process of optimizing the model's objective function. *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 35, at 79–80.

[41] *See* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7, at 25 ("Unlocking the full potential of machine learning in any regulatory context, but especially in the enforcement context, requires abundant, well-labeled data that accurately reflect 'ground truth' about misconduct."); RUSSELL & NORVIG, *supra* note 29, at 653–56 (discussing supervised learning).

[42] *See* Alexander Campolo & Kate Crawford, *Enchanted Determinism: Power Without Responsibility in Artificial Intelligence*, 6 ENGAGING SCI. TECH. & SOC'Y. 1, 10–12 (2020) (suggesting that machine learning ultimately boils down to "detecting complex patterns in nonlinear ways from large data sets").

B. *Bias In, Bias Out*[43]

Although some suggest that machine learning is en route to fundamentally alter the structure of knowledge itself, use of this technology to assist human decision-making is still in its relative infancy.[44] Because we are still struggling to understand the full scope of its implications, we are at risk of falling prey to "local maxima" by fixating on glaring failures and failing to observe more fundamental risks that lurk just around the corner.[45] Two such failures currently dominate the discussion: machine learning's tendency to produce hidden biases and its opacity.[46] In both cases, the design features that give rise to these challenges result from the "input" side of the algorithm's design; their dominance can draw attention away from "output" concerns, such as those surrounding the algorithm's measure of relevance and the resulting effects of the filter bubble.[47]

Talk of an algorithm's hidden biases commonly refers to the training process's propensity to overemphasize the weight of input variables directly or indirectly tied to protected personal characteristics such that the result would be a discriminatory algorithm.[48] Such failures can result from the infamous "garbage in, garbage out" problem that

[43] *See* Mayson, *supra* note 11.

[44] MAYER-SCHÖNBERGER & CUKIER, *supra* note 32, at 5 (arguing that big data and the use of machine learning to process it represent a fundamental shift in the structure of knowledge).

[45] I thank Josh Fairfield for suggesting this framing.

[46] *See, e.g.*, Bamberger, *supra* note 7, at 723–24 (discussing the challenge of algorithmic opacity); Calo, *supra* note 11, at 415 (discussing the scholarly concern with biases and the lack of transparency); Lehr & Ohm, *supra* note 28, at 668 (describing opacity and lack of explainability as "some of the most viscerally unsettling harms of machine learning"); Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1, 2 (2018) (identifying opacity and algorithms' discriminatory effects as two of the main problems dominating scholarship); Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018) (discussing the problem of explainability); John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 PHIL. & TECH. 245 (2016) (discussing opacity and "hiddenness").

[47] *See* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 143–44 (2018) (discussing the choice of output variable with regard to child maltreatment prediction systems); Lehr & Ohm, *supra* note 28, at 675 (discussing the importance of the output variable).

[48] *See* Barocas & Selbst, *supra* note 11 (discussing algorithms' disparate effects); Ben Green & Yiling Chen, *Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts*, PROC. ACM ON HUM.–COMPUT. INTERACTION, Oct. 2021, at 1, 3 (discussing how algorithms can exacerbate racial disparities); Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 U.C. DAVIS L. REV. 1067, 1075 (2018) (discussing how algorithms can produce results that disproportionately harm minorities); *see also* Sebastian Benthall & Bruce D. Haynes, *Racial Categories in Machine Learning*, 2019 ASS'N COMPUTING MACH. 289.

plagues machine learning.[49] As this adage suggests, any significant flaws or deficiencies in the training data will inevitably, unless cleaned or compensated, find their way into the operating algorithm.[50] When algorithms are trained on datasets tainted by bias and discrimination, the resulting algorithm will come out similarly biased.[51] Such distortions can develop even when the examples in the datasets are not themselves directly biased. As with any design process, the creation of a learning algorithm is constrained by cost-effectiveness and the availability of relevant training data.[52] As Deirdre Mulligan and Kenneth Bamberger demonstrate, the decisions such restrictions force on developers are rarely value-neutral.[53] Tilted design choices, Bamberger shows, can distort the ensuing model and produce a similarly skewed algorithm.[54]

Algorithmic biases can be challenging to detect. The gold standard for evaluating an algorithm's accuracy is to test it on "unseen data," meaning data that was not included in the datasets on which it trained.[55] However, the data in unseen data tests are commonly taken from the same source that produced the training data.[56] As a result, any hidden biases in the training process would not necessarily hinder the algorithm's ability to accurately perform in testing.[57] Even when the

[49] *See* KELLEHER & TIERNEY, *supra* note 33, at 47–48 (discussing the "garbage in, garbage out" problem).

[50] *See id.* at 35–36 (discussing how bad data collection can lead to bad results); Ferguson, *supra* note 7, at 401–02 (discussing how biased data can produce biased algorithmic predictions); Kroll et al., *supra* note 6, at 681 (discussing how machine-learning models can build in discrimination).

[51] For discussions on algorithmic biases, see Barocas & Selbst, *supra* note 11; Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Mayson, *supra* note 11.

[52] *See* Lehr & Ohm, *supra* note 28, at 675 (discussing the practical considerations that shape the algorithm's design).

[53] Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-by-Design*, 106 CALIF. L. REV. 697, 718 (2018) (discussing the algorithmic distortions that "arise from the social and technical environment in which regulatory norms are 'translated' into hardwired code and include the cognitive biases of those who design and use the technologies" (footnotes omitted)).

[54] Bamberger, *supra* note 7, at 728 ("The need to translate both legal and management concerns into a third distinct logic of computer code and quantitative analytics creates the possibility that legal choices will be skewed by the biases inherent in that process.").

[55] *See* KELLEHER, *supra* note 30, at 14 (discussing unseen data testing).

[56] *See* Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla & Francisco Herrera, *A Unifying View on Dataset Shift in Classification*, 45 PATTERN RECOGNITION 521, 526–28 (2011) (discussing sample-selection bias and the challenge of correcting it).

[57] *See* Alexander D'Amour et al., *Underspecification Presents Challenges for Credibility in Modern Machine Learning*, ARXIV, Nov. 24, 2020, at 2–3, https://arxiv.org/pdf/2011.03395.pdf [https://perma.cc/KYH2-RNB5] (suggesting that such testing procedures "are agnostic to the particular inductive biases encoded by the trained model," and suggesting that this challenge is further compounded by "underspecification").

algorithm begins operating in the real world, it can be challenging to spot its hidden biases since the algorithm's accuracy will often be measured against the same flawed data that produced it.[58] Moreover, as results of the algorithm's operation are fed back into the training data for retraining, biases can produce a "runaway feedback" effect as the algorithm becomes progressively more discriminatory.[59]

Even when biases are detected, it can be challenging to purge them out of the model animating the algorithm. Due to their ability to create highly complex models, advanced forms of machine learning are susceptible to "overfitting" the model to the training data, incorporating noise patterns that happen to correlate with the desired function.[60] Hence, even when input data relating to protected classifications are omitted from the training sets, overfitting can overemphasize the weight of noise patterns these datapoints leave in their wake, potentially producing a model that is just as biased.[61]

These failures connect to the second challenge characteristic of machine learning, namely its notorious opacity, likewise the subject of much critical attention.[62] Although ensuring an algorithm's accuracy and impartiality requires knowledge of its operation, machine learning can frustratingly pit accuracy and transparency against each other.[63] State-of-the-art learning technologies have been highly successful at emulating human decision-making by tapping into the multifaceted and intuitive depths of human reasoning; these breakthroughs, however, often come at the expense of explainability, as they produce

[58] *See, e.g.*, Tammy Wang, *How Machine Learning Will Shape the Future of Hiring*, LINKEDIN: PULSE (Mar. 8, 2017), https://www.linkedin.com/pulse/how-machine-learning-shape-future-hiring-tammy-wang [https://perma.cc/MS3Q-E7D6] (describing the complexity of measuring successful hiring decisions).

[59] *See* Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger & Suresh Venkatasubramanian, *Runaway Feedback Loops in Predictive Policing*, ARXIV, Dec. 22, 2017, https://arxiv.org/pdf/1706.09847.pdf [https://perma.cc/2BRL-423U] (describing the occurrence of runaway feedback loops).

[60] On overfitting, see KELLEHER, *supra* note 30, at 21. On how it can create disparate effects, see Kroll et al., *supra* note 6, at 681; Lehr & Ohm, *supra* note 28, at 704.

[61] On the pervasiveness of biases, see Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 28 (2017); Kroll et al., *supra* note 6, at 681.

[62] For scholarship on the problem of opacity, see Joshua P. Davis, *Law Without Mind: AI, Ethics, and Jurisprudence*, 55 CAL. W. L. REV. 165, 181–82 (2018); Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 4 (2017); Frank Pasquale & Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviourism*, 68 U. TORONTO L.J. 63, 63 (2018); Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 184 (2017).

[63] *See* Selbst & Barocas, *supra* note 46, at 1088 (discussing explainability).

so-called computational "black boxes."[64] To achieve accuracy in such tasks, machine learning creates hyperdimensional models that are often impervious to human comprehension in their complexity.[65] Furthermore, with the use of "deep" learning, and in particular convolutional neural networks, such complex models can include variables in the form of mathematical abstractions that are not explicitly found in the input data and can be devoid of comprehensible semantic meaning.[66]

Even when the algorithms themselves involve no mathematical complexity, opacity can result from trade secrecy, similarly limiting the ability to scrutinize the algorithm's accuracy and impartiality.[67] Although many simple algorithms are precise mathematical formulae, ironically, their relative transparency makes them secretive: As precise descriptions of their own operation, such algorithms are, in a sense, themselves the technology they embody.[68] Since most algorithms are created by private companies even when they are used for public purposes, revealing the details of the algorithm would essentially deprive these companies of their trade secrets.[69] Likewise, since algorithms are descriptions of their operation, keeping them secret is at times necessary to prevent gaming.[70]

Both characteristic failures are worthy of the critical attention they receive. Biased decision-making can have a pervasive, corrupting, and delegitimizing effect, and opacity can be just as damning, making it impossible to discern whether the algorithm impartially evaluates different datapoints. Still, focusing on the actual or potential biased

[64] On machine learning computational black boxes, see J.M. Benítez, J.L. Castro & I. Requena, *Are Artificial Neural Networks Black Boxes?*, 8 IEEE TRANSACTIONS ON NEURAL NETWORKS 1156, 1156–57 (1997).

[65] *Id.*

[66] *See* KELLEHER, *supra* note 30, at 129–43 (providing a succinct explanation of the ability of neural networks to abstract away from readily available representations of the data). On the effect of using convolutional neural networks, see Shlomit Yanisky Ravid & Xiaoqiong (Jackie) Liu, *When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3A Era*, 39 CARDOZO L. REV. 2215, 2226 (2018).

[67] On the connection between trade secrecy and algorithmic opacity, see Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54 (2019).

[68] For this description of algorithms, see Kroll et al., *supra* note 6, at 646.

[69] *See* Simmons, *supra* note 48, at 1087 ("Two of the largest providers of predictive algorithms in the criminal justice system are corporations who claim that the inner workings of their software are trade secrets."); Perel & Elkin-Koren, *supra* note 62, at 184–85 ("[A]lgorithmic decision-making is essentially concealed behind a veil of a code, which is often protected under trade secrecy law . . . .").

[70] On the problem of gaming and secrecy, see Bambauer & Zarsky, *supra* note 46; Ignacio N. Cofone & Katherine J. Strandburg, *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. 623 (2019).

weighing of input variables can come to mean that little attention is devoted to the *output* variable, meaning the classification that the algorithm seeks to reproduce.

C. *Errors and Urgency*

In cases like the use of algorithms to prevent nonconsensual pornography, neglect of long-term systemic effects such as those of the filter bubble can also be explained by the urgency of the harms the algorithm is meant to prevent. Considering these harms' immediacy and gravity, the failures that seem to require the most immediate attention are those that prevent the algorithm from properly addressing these harms or that produce comparable harms, such as discriminatory decision-making. As one commentator suggested in the similar setting of flagging child maltreatment, when such grave harms are on the line, "[i]t is hard to conceive of an ethical argument against use of the most accurate predictive instrument."[71] Or, as Ryan Calo puts it, "As the saying goes, 'justice delayed is justice denied': we should not aim as a society to hold a perfectly fair, accountable, and transparent process for only a handful of people a year."[72]

This sense of urgency can become evident in the choice of the output variable, as limited time and resources often force designers to use variables found in readily available datasets, even when these output variables do not perfectly match the desired function.[73] Using such proxies inevitably means that the algorithm performs a function that is not identical to the task it is thought to be performing.[74] In the case of child maltreatment, for instance, such considerations have led designers to use the decision to place a child in foster care as a stand-in for maltreatment—despite the incongruence of the two and with the full awareness that an agency's decision is not necessarily indicative of harm—simply because no other reliable data was available.[75] The result of such substitution is an algorithm that does not determine the

[71] Dan Hurley, *Can an Algorithm Tell When Kids Are in Danger?*, N.Y. TIMES (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html [https://perma.cc/ER5B-LYUP] (quoting Marc Cherna, the director of Allegheny County's Department of Human Services).

[72] Calo, *supra* note 11, at 415.

[73] *See* Lehr & Ohm, *supra* note 28, at 675 (discussing how practical considerations can affect the choice of output variable).

[74] *Id.* ("[P]ursuing a particular outcome variable for the sake of convenience carries with it a greater risk of mismatch between the predictive goal and the variable's specification.").

[75] *See* EUBANKS, *supra* note 47, at 143–44. An additional proxy was a repeated referral in cases that did not initially involve an agency response.

probability that a child is at risk but instead determines the likelihood that a human decision-maker would find him or her to be at risk.

Modeling available data rather than ideal training data can considerably distort the algorithm's operation as a result of selection bias, meaning the training data's failure to be adequately representative of the real world.[76] In the typical example of credit scoring, an algorithm can be trained on a database that includes a large number of loan applications with relatively few minority applicants, making resulting predictions less accurate for future minority applications.[77] Thus, even when the individual decisions in the datasets are unbiased and the training algorithm correctly models them, the running model will fail to impartially decide on the likelihood of loan default.[78] In such cases, it can be said that the algorithm attained an inaccurate "concept" of default risk.[79] Similarly, the attempt to capture an accurate concept of nonconsent could be distorted by turning to readily available decisions made in response to takedown requests or court opinions made in criminal procedures.[80] A concept of consent attained from takedown decisions will inevitably be geared toward complainants who were sufficiently informed about such procedures and had the means and social capital needed to advance their cause;[81] a concept of consent attained from judicial decisions will be shaped both by those

[76] *See* D'Amour et al., *supra* note 57, at 2 ("[A] predictor trained in a setting that is structurally misaligned with the application will reflect this mismatch.").

[77] *See* Lehr & Ohm, *supra* note 28, at 680–81 (discussing this example).

[78] *Id.* at 680 (discussing the emergence of this bias).

[79] On "concept attainment," see Jeffrey C. Schlimmer & Richard H. Granger, Jr., *Incremental Learning from Noisy Data*, 1 MACH. LEARNING 317, 317–18 (1986); Gerhard Widmer & Miroslav Kubat, *Learning in the Presence of Concept Drift and Hidden Contexts*, 23 MACH. LEARNING 69 (1996). On its meaning for decision-making algorithms, see Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 140–42 (2017); Kroll et al., *supra* note 6, at 680.

[80] *See* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7, at 25.

[81] On the selection bias in takedown requests, see Nick Hopkins & Olivia Solon, *Facebook Flooded with "Sextortion" and "Revenge Porn," Files Reveal*, GUARDIAN (May 22, 2017, 9:52 AM), https://www.theguardian.com/news/2017/may/22/facebook-flooded-with-sextortion-and-revenge-porn-files-reveal [https://perma.cc/AVJ9-NPNN]; Sarah Bloom, Note, *No Vengeance for "Revenge Porn" Victims: Unraveling Why This Latest Female-Centric, Intimate-Partner Offense Is Still Legal, and Why We Should Criminalize It*, 42 FORDHAM URB. L.J. 233, 253–54 (2014); Alisha Kinlaw, Comment, *A Snapshot of Justice: Carving Out a Space for Revenge Porn Victims Within the Criminal Justice System*, 91 TEMP. L. REV. 407, 421–29 (2019). On Facebook's problematic history with takedown requests, see Solon, *supra* note 36; Alexandra Topping, *Facebook Revenge Pornography Trial "Could Open Floodgates,"* GUARDIAN (Oct. 9, 2016, 10:21 AM), https://www.theguardian.com/technology/2016/oct/09/facebook-revenge-pornography-case-could-open-floodgates [https://perma.cc/MT58-QKY5].

considerations as well as by selection biases that result from police officers' and prosecutors' preferences.[82]

It is important to realize that despite the facial similarities between conceptual distortions and the challenge of legalistic filter bubbles, they represent two different understandings of failure. Conceptual failures are essentially failures to correctly translate the task the algorithm is meant to perform into a suitably labeled function, commonly due to the unavailability of better-suited training data. In the case of algorithms meant to perform legal tasks, this failure manifests in the misconstruction of the relevant legal paradigm. In contrast, the effect of legalistic filter bubbles goes beyond such errors to the adverse effects filtering can have even when it correctly implements the legal norm—even when, for instance, it implements a conception of consent that is entirely congruent with its legal meaning. This brings us to the inevitable question—What could be wrong with the correct implementation of a legal norm?

D. *Taking Law (Too) Seriously*

Naturally, algorithms that use legal analysis to determine information's relevance are commonly used to assist in law or law-adjacent undertakings. In such normative settings, reliance on algorithmic filtering can produce what Tim Wu describes as human-machine hybrid social-ordering ecosystems.[83] Although these normative ecosystems are not always formally part of the legal process, their reliance on legal norms and influence on these norms' development can be substantial.[84] Decisions made in these systems are often informed by controlling legal norms that proscribe unfair, discriminatory, or otherwise illicit considerations or results; at times, as with nonconsensual pornography, controlling legal norms also inform

[82] Despite significant reforms aiming to align sexual assault laws with an expansive understanding of sexual autonomy, police officers still commonly fail to investigate, and prosecutors refrain from charging, cases that do not involve physical force. *See* Donald Dripps, *After Rape Law: Will the Turn to Consent Normalize the Prosecution of Sexual Assault?*, 41 AKRON L. REV. 957, 975 (2008); Deborah Tuerkheimer, *Rape On and Off Campus*, 65 EMORY L.J. 1, 4 (2015); Patricia J. Falk, *Husbands Who Drug and Rape Their Wives: The Injustice of the Marital Exemption in Ohio's Sexual Offenses*, 36 WOMEN'S RTS. L. REP. 265, 286 (2015). For a general discussion of the processing of sexual assault reports, see Cassia Spohn, Clair White & Katharine Tellis, *Unfounding Sexual Assault: Examining the Decision to Unfound and Identifying False Reports*, 48 LAW & SOC'Y REV. 161 (2014).

[83] Wu, *supra* note 7.

[84] *See* Perel & Elkin-Koren, *supra* note 62, at 191 ("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.").

the direction of decision-making.[85] Furthermore, decisions made in these hybrid systems can set the tone for subsequent legal adjudication regardless of whether they are formally part of legal adjudication or merely adjacent to it.[86]

As a result, much of the existing criticism of algorithmic decision-making has focused on its propensity for transgressing or distorting the legal norms that control the algorithm's operation.[87] The commonly suggested cure to the legitimacy deficits such algorithmic failures produce is a demand for significant human involvement "in the loop" of the algorithm's creation and operation.[88] Demands for human involvement have included calls to incentivize algorithm-design insiders to report on any illicit elements, demands for a meaningful role for human decision-makers, insistence on regulatory oversight, and suggestions that outside researchers be provided access to the algorithm.[89] Generally, the purpose of greater human involvement is to ensure that the algorithm correctly implements the norms that guide its operation and does not transgress any general norms that constrain it, such as the legal prohibitions on disparate treatment.[90]

However, this presumed division of labor between the algorithm and the human decision-maker risks oversimplifying the relationship between the two components of the human-machine hybrid ecosystem.

[85] On the interaction between private algorithmic adjudication and law enforcement, see Douek, *supra* note 11; Niva Elkin-Koren & Eldar Haber, *Governance by Proxy: Cyber Challenges to Civil Liberties*, 82 BROOK. L. REV. 105 (2016).

[86] *See, e.g.*, Bamberger, *supra* note 7, at 676; Calo, *supra* note 11, at 415; Elkin-Koren, *supra* note 7, at 2.

[87] For emphasis on the algorithm's accurate implementation of the law, see Calo, *supra* note 11, at 415 ("[A]ccuracy is an important dimension of fairness."); Kroll et al., *supra* note 6, at 681–82 ("[T]ransparency and after-the-fact auditing can only go so far in preventing undesired results."); Simmons, *supra* note 48, at 1070 ("While many commentators argue that predictive algorithms pose a severe threat to the fairness of the criminal justice system, these tools will increase the accuracy, efficiency, and fairness of many aspects of policing and adjudication if instituted properly." (footnote omitted)).

[88] *See* Lehr & Ohm, *supra* note 28, at 657 (discussing the human-in-the-loop approach); *see also* Kiel Brennan-Marquez & Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137, 146–48 (2019); Calo, *supra* note 11, at 415–16; Desai & Kroll, *supra* note 62, at 51; Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 6 (2019); Rich, *supra* note 7, at 898.

[89] *See* Bamberger, *supra* note 7, at 729, 736–38 (suggesting more regulatory oversight and greater human involvement and oversight); Katyal, *supra* note 67, at 130–37 (suggesting whistleblower protection to encourage transparency); Perel & Elkin-Koren, *supra* note 62, at 185–86 (suggesting access to external researchers).

[90] *See* Perel & Elkin-Koren, *supra* note 62, at 198 (suggesting ways to interact with the algorithm's black box to prevent errors); Lehr & Ohm, *supra* note 28, at 704–05 (discussing ways of addressing discrimination); Simmons, *supra* note 48, at 1075, 1101 ("[P]redictive algorithms give us an opportunity to overcome bias because we can monitor the data that the algorithms use and, if need be, correct for pre-existing racial discrimination.").

To begin with, talk of *introducing* human agency into the algorithm's creation and operation disregards the fact that, for the most part, these processes are already shot through with human involvement.[91] Translating any task into a programmable undertaking, assembling and structuring the datasets, and designing the learning algorithm are inescapably human endeavors that significantly affect the working algorithm's operation.[92] Therefore, all too often, talk of "introducing" a human into the algorithm's ecosystem can simply come to mean designating some human actors, out of the many already involved in the process, as accountability lightning rods.[93]

Seeing the problem as one concerning the limits to the algorithm's involvement further obscures structural issues that persist even when the algorithm performs only an assistive function. As several authors note, the use of algorithms can have considerable effects on the meaning of legal norms even when human decision-makers are in complete control of the decision itself.[94] Ben Green and Yiling Chen, for instance, note how the use of risk-assessment algorithms can increase the salience of risk in the decision, deviating from the balance set by applicable legal norms.[95] Cary Coglianese and David Lehr discuss a similar algorithm-induced shift toward reliance on quantitative judgments.[96] Richard Re and Alicia Solow-Niederman likewise argue that the efficiency of algorithmic adjudication can inspire a turn toward "codified justice," meaning an interpretation of legal norms that favors standardization over judicial discretion.[97] Finally, Andrew Ferguson notes that reliance on the products of algorithmic data analysis can lead users to trust in their worst instincts as the algorithm presents them with lopsided results.[98]

The common thread that runs through these insightful criticisms is the idea that for legal and law-adjacent tasks, the main downside to relying on algorithmic assistance is that doing so might produce legally

[91] *See* Lehr & Ohm, *supra* note 28, at 657 (suggesting that the call to introduce a human in-the-loop could be based on a conflation between the algorithm's creation and its operation).

[92] *See id.* at 660 (discussing the various human involvements in the algorithm's creation).

[93] *See* Calo, *supra* note 11, at 416 (discussing the concern that soldiers used alongside automated weapons systems "will be placed into the loop for the sole purpose of absorbing liability for wrongdoing").

[94] *See* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1218 (2017) (discussing the effects the use of algorithms can have on decision-makers).

[95] Green & Chen, *supra* note 48, at 1.

[96] Coglianese & Lehr, *supra* note 94.

[97] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 246 (2019).

[98] *See* Ferguson, *supra* note 7, at 402.

incorrect results or otherwise distort the meaning of the applicable legal regime. Some blame this failure on the intricate interaction between law and social reality, which makes it difficult, if not impossible, for algorithms to accurately emulate legal analysis. Frank Pasquale, for instance, suggests that law's social embeddedness makes it likely that legal algorithms will fail to capture the "particular political systems and traditions" that govern the legal norm, so use of these algorithms "is unlikely to meet the complex standards of review and appeal embodied in the Legal Process conception of the rule of law."[99] Harry Surden similarly writes that "in many instances in legal prediction there may be subtle factors that are highly relevant to legal prediction and that attorneys routinely employ in their professional assessments," arguing that these factors may be difficult for machine-learning algorithms to attain.[100] A similar view is put forward by Joshua Kroll et al., who raise concerns about algorithms' failure to live up to the human ability to interpretively fill in details intentionally left vague by legislatures.[101]

These criticisms, however, confine themselves to the algorithm's ability to accurately implement the law, understood as the sum of requirements made by all applicable legal rules and standards. By doing so, these criticisms neglect the critical need for legal adjudication to look *beyond* current law to the legally immaterial considerations that can shape the law's development.[102] For the problem caused by legalistic filter bubbles is not that it distorts law, but rather that it prevents it from evolving.

II. LAW IN, LAW OUT

On October 15, 2017, actor Alyssa Milano used her Twitter account to urge people to share their experiences of sexual wrongdoing.[103] Quoting a message that participated in the "Me Too" movement,[104] Milano added, "If you've been sexually harassed or

[99] Pasquale, *supra* note 88, at 45.

[100] Surden, *Machine Learning*, *supra* note 6, at 106.

[101] Kroll et al., *supra* note 6, at 679.

[102] *See* Elkin-Koren, *supra* note 7, at 7 ("[A]s we shift to governance of speech by ML, we are losing an important opportunity to scrutinize norms and socially negotiate the values tradeoffs they embed.").

[103] *See* Alyssa Milano (@Alyssa_Milano), TWITTER (Oct. 15, 2017, 4:21 PM), https://twitter.com/Alyssa_Milano/status/919659438700670976 [https://perma.cc/YHR9-2R4L].

[104] On the origins of #MeToo, see Sandra E. Garcia, *The Woman Who Created #MeToo Long Before Hashtags*, N.Y. TIMES (Oct. 20, 2017), https://www.nytimes.com/2017/10/20/us/me-too-movement-tarana-burke.html [https://perma.cc/MQ5P-JEG7]; Abby Ohlheiser, *The Woman*

assaulted write 'me too' as a reply to this tweet."[105] The tweet went viral, eliciting tens of thousands of responses that shared personal experiences of sexual violation. Many of these responses, however, as well as many of those that were later discussed in the context of the #MeToo movement, as it became known after the tweet, did not involve experiences that could be categorized as sexual harassment or assault, at least not in their legal meaning.[106] Often these stories portrayed incidents that did not involve workplace discrimination and that were not nonconsensual in the legal sense of the term.[107] Although some have criticized the #MeToo movement for this transgression of dominant legal norms, I have argued elsewhere that this departure is better understood as a call to acknowledge current law's limited ability to capture the wrongness of consensual but undesirable, exploitative, and demeaning sexual interactions.[108] On this view, the #MeToo movement has had such a profound impact also *because* it challenged consent's dominance in the discussion of sexual wrongdoing and confronted legal thinking with all the painful experiences that lie beyond the law's reach.[109]

I will return to the question of consent in greater detail in Part III. In the following pages, I will suggest that if Twitter's algorithm were to measure responses' relevance according to their legal classification instead of the direct connection responders made between their own personal stories and Milano's tweet, this opportunity for normative evolution would fail to materialize. I will refer to this phenomenon as the emergence of a *legalistic* filter bubble, suggesting that it can occur when algorithms replace human beings in determining the normative relevance of information according to its legal classification.

A. *The Filter Bubble*

"It is some time in the future. Technology has greatly increased people's ability to 'filter' what they want to read, see, and hear."[110]

Behind "Me Too" Knew the Power of the Phrase When She Created It—10 Years Ago, WASH. POST (Oct. 19, 2017), https://www.washingtonpost.com/news/the-intersect/wp/2017/10/19/the-woman-behind-me-too-knew-the-power-of-the-phrase-when-she-created-it-10-years-ago [https://perma.cc/2M3M-EHQN].

[105] *See* Milano, *supra* note 103.

[106] *See* Maggen, *supra* note 24, at 598–605 (discussing #MeToo's relation to sexual assault and harassment).

[107] *See id.*

[108] *See id.* at 610–15.

[109] *See id.* at 615–34.

[110] CASS R. SUNSTEIN, REPUBLIC.COM 2.0, at 1 (2009).

Writing this in 2009, Cass Sunstein predicted that technology would allow people to choose only to view the information that fits their interests, "no more and no less."[111] "When the power to filter is unlimited," Sunstein writes, "people can decide, in advance and with perfect accuracy, what they will and will not encounter."[112] If unlimited filtering gains hold over most people's information-consumption habits, Sunstein predicts, this would have devastating effects on democratic governance and society's political functioning, as it would erode the range of commonly shared experiences and reduce people's chances of encountering opposing worldviews.[113]

As Sunstein later came to realize, technological changes can make filtering much less about consumers' ability to *choose* what they see and more about the design choices informing the filtering mechanisms and the business models that drive these choices.[114] In 2010, Eli Pariser coined the term "filter bubble" to express this notion, highlighting how the design of filtering algorithms can constrict users' worldviews regardless of their choice in the matter.[115] In the filter bubble metaphor, the shape and color of the filter, so to speak, are not directly determined by the user's choices but rather by the design of the algorithms that act as buffers between the user and the world.[116] In this scheme, the main design feature determining the filtering's effect is the algorithm's method for determining the *relevance* of available information.[117] Naturally, relevance is a broad and malleable term, but generally speaking, the purpose of any filtering algorithm is to provide the best results, and this commonly translates to those results that are most relevant to its presumed purpose.[118] The concepts that translate accuracy into the specific measures of relevance thus effectively determine what users see and, ultimately, how they view the world.

[111] *Id.*

[112] *Id.* at 3.

[113] *See id.* at 5–6.

[114] *See generally* CASS R. SUNSTEIN, CHOOSING NOT TO CHOOSE: UNDERSTANDING THE VALUE OF CHOICE 34–35 (2015); CASS R. SUNSTEIN, #REPUBLIC: DIVIDED DEMOCRACY IN THE AGE OF SOCIAL MEDIA (2017) [hereinafter SUNSTEIN, #REPUBLIC].

[115] PARISER, *supra* note 9.

[116] *See id.* at 30–32.

[117] *Id.* at 21–46; *see also* Gillespie, *supra* note 10.

[118] As Gillespie notes, however,

> As there is no independent metric for what *actually* are the most relevant search results for any given query, engineers must decide what results look "right" and tweak their algorithm to attain that result, or make changes based on evidence from their users, treating quick clicks and no follow-up searches as an approximation, not of relevance exactly, but of satisfaction.

Gillespie, *supra* note 10, at 175.

Pariser's account of the filter bubble revolves around Google's 2009 decision to expand the personalization of its search engine to include all users, not just those logged in to their Google accounts, as had been the case since 2005.[119] This seemingly innocuous design choice, Pariser notes, reflected a profound shift in the search algorithm's objective that was in turn reflected in how it measures websites' relevance to a given query.[120] Before that change, the purpose of the algorithm was to locate the websites most relevant to the query's search terms; since the change, the purpose of the algorithm has been to produce the results most relevant to the *user's* query, as inferred from the information Google collects about the user from various sources.[121] Consequently, different users will be provided different results for the same query because of the algorithm's assessments of their diverging personal preferences.[122]

It is worth taking stock of the leap from Sunstein's foreboding predictions to the realization that filter bubbles are *already* here to stay. A significant aspect of this shift is recognizing that some form of filtering is an inevitable feature of the information age; personalization and the polarization that supposedly ensues from personalized filtering are private instances of the broader filtering dynamic.[123] What makes filter bubbles so potent is that filtering can be an unavoidable necessity when dealing with otherwise prohibitively vast amounts of data.[124] In most cases, unless users are interested in specific information and know exactly where to find it, the digital information they obtain through their use of search engines, recommendation algorithms, catalogs, and other querying mechanisms will come pre-filtered to show them only those bits of information that are presumed to be relevant to their needs, with users only minimally conscious of the hidden selection involved.[125] The digital age's "information overflow" essentially turns the prospect of perfect filtering on its head: instead of serving solipsist users' desire

[119] *See* Horling & Kulick, *supra* note 14.

[120] PARISER, *supra* note 9, at 30–36.

[121] *Id.* at 1–3.

[122] *See* Horling & Kulick, *supra* note 14.

[123] As Marshall McLuhan suggests, this may be true of any technology. MARSHALL MCLUHAN, UNDERSTANDING MEDIA: THE EXTENSIONS OF MAN 8, 56 (1964).

[124] *See* Deven R. Desai, *Exploration and Exploitation: An Essay on (Machine) Learning, Algorithms, and Information Provision*, 47 LOY. U. CHI. L.J. 541, 549 (2015) ("Search, social networks, online rating systems, tweets, apps, and mobile computing have emerged to aid us as we try to make sense of the world, and these advances generate the perceived problem of perfect filtering, echo chambers, and walled gardens.").

[125] PARISER, *supra* note 9, at 82 ("Like a lens, the filter bubble invisibly transforms the world we experience by controlling what we see and don't see."); Gillespie, *supra* note 10, at 183 ("[A]lgorithms impinge on how people seek information, how they perceive and think about the contours of knowledge, and how they understand themselves in and through public discourse.").

to insulate themselves from undesirable information, it becomes indispensable to the ability to engage with a world of endless riches of information.[126] What we should ask, therefore, is not whether effective filtering is possible, but rather what the effects of the filtering already in place are. As Tarleton Gillespie puts it, "This means we must consider not [algorithms'] 'effect' on people, but a multidimensional 'entanglement' between algorithms put into practice and the social tactics of users who take them up."[127]

Although it centers on the presumed consequences of personalized filtering, the filter bubble theory can be read as more generally suggesting that the winnowing effect caused by filtering is prone to entrenching users' acceptance of the applied measures of relevance and suppressing information that could undermine it.[128] In this sense, using the user's personal preferences as the applicable measure of relevance can entrench the user's tendency to take these preferences for granted and diminish the user's opportunity to encounter information that conflicts with them.[129] However, similar effects can occur with different choices of relevance.[130] Personalization, in this sense, simply made filtering's effects more apparent as it coincided with familiar scholarly themes that, as early as the 1990s, warned of the perils of personalized news consumption.[131] However, algorithmic filtering has similar effects with other measures of relevance as well, causing comparable harms.[132]

The contours of this idea can be best gleaned from the critical responses its warnings elicited. In the specific setting of personalized algorithms, the filter bubble theory suggests that the ills of filtering are twofold. First, from a political perspective, personalized information consumption can breed political destabilization, radicalization, and polarization as people miss out on opportunities to engage with

[126] PARISER, *supra* note 9, at 22 ("The solution to the information overflow of the digital age was smart, personalized, embedded editors."); Desai, *supra* note 124, at 547 ("Algorithms, public experts, social networks, online rating systems, and more have emerged to help us as we again try to sort information overload.").

[127] Gillespie, *supra* note 10, at 183.

[128] PARISER, *supra* note 9, at 84 ("First, the filter bubble surrounds us with ideas with which we're already familiar (and already agree), making us overconfident in our mental frameworks. Second, it removes from our environment some of the key prompts that make us want to learn.").

[129] *Id.* at 15 ("In the filter bubble, there's less room for the chance encounters that bring insight and learning.").

[130] Gillespie, *supra* note 10, at 187 ("[A]lgorithms designed to offer relevant knowledge also offer ways of knowing—and . . . as they become more pervasive and trusted, their logics are self-affirming.").

[131] *See* NICHOLAS NEGROPONTE, BEING DIGITAL 1–18, 153 (1995) (discussing *The Daily Me*, a hypothetical personalized newspaper).

[132] *See generally* Gillespie, *supra* note 10 (discussing the effect of algorithms' measure of relevance).

opposing thoughts and values and as the repository of shared communal experiences vital to democratic governance is depleted.[133] Second, from a consumer perspective, personalized filtering involves a growing information asymmetry as tech companies gain greater insight into commercially exploitable user preferences.[134]

Although intuitively compelling, these arguments have recently faced considerable criticism.[135] The rebuttals, however, mainly concern the specifics of personalized filtering and do not necessarily undermine the more general aspects of the theory. Empirical studies have purported to show that personalized filtering does not produce political polarization and radicalization, or that these apprehensions have been overstated.[136] Such evidence, however, has only limited bearing on the general applicability of the filter bubble theory, especially as it concerns the effects of legal filtering.

Other critics have more broadly suggested that personalized filter bubbles fail to produce observable political polarization because perfect filtering is impossible; without it, they suggest, filtering's ostensible effects are minimal.[137] Elizabeth Dubois and Grant Blank argue that "[w]hatever may be happening on any single social media platform, when we look at the entire media environment, there is little apparent echo chamber."[138] Nevertheless, the fact that algorithmic filtering cannot fully isolate most of its users does not mean that it cannot dominate confined avenues where algorithms act as gatekeepers by

[133] On these effects, see CASS R. SUNSTEIN, GOING TO EXTREMES: HOW LIKE MINDS UNITE AND DIVIDE 23–25 (2009); Gerhard Wagner & Horst Eidenmüller, *Down by Algorithms? Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions*, 86 U. CHI. L. REV. 581, 598 (2019).

[134] PARISER, *supra* note 9, at 146–47 (discussing the relation between filter bubbles and knowledge asymmetry); *see also* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (2019).

[135] For such criticisms see, for example, Eytan Bakshy, Solomon Messing & Lada A. Adamic, *Exposure to Ideologically Diverse News and Opinion on Facebook*, 348 SCIENCE 1130 (2015); Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker & Richard Bonneau, *Tweeting from Left to Right: Is Online Political Communication More than an Echo Chamber?*, 26 PSYCH. SCI. 1531 (2015); Elizabeth Dubois & Grant Blank, *The Echo Chamber Is Overstated: The Moderating Effect of Political Interest and Diverse Media*, 21 INFO. COMMC'N & SOC'Y 729, 740 (2018); Seth Flaxman, Sharad Goel & Justin M. Rao, *Filter Bubbles, Echo Chambers, and Online News Consumption*, 80 PUB. OP. Q. 298 (2016); Jessica T. Feezell, *Agenda Setting Through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era*, 71 POL. RSCH. Q. 482 (2018); Dan Hunter, *Philipic.com*, 90 CALIF. L. REV. 611 (2002) (reviewing CASS R. SUNSTEIN, REPUBLIC.COM (2001)).

[136] For such studies, see, for example, Barberá, Jost, Nagler, Tucker & Bonneau, *supra* note 135, at 1531–32; Flaxman, Goel & Rao, *supra* note 135, at 299; Feezell, *supra* note 135.

[137] For this argument, see, for example, Bakshy, Messing & Adamic, *supra* note 135, at 1130–31; Hunter, *supra* note 135, at 614.

[138] Dubois & Blank, *supra* note 135, at 740.

controlling information bottlenecks. This can be the case with algorithms used for legal or law-adjacent purposes. Admittedly, in most cases, law enforcement relies on information obtained from a variety of sources, only some of which currently involve algorithmic filtering. Still, in some cases, when investigations require screening massive amounts of data for "victimless" crimes or violations, or when those adversely affected seldom complain, screening publicly available data using algorithmic filtering can become the legal discourse's primary source of information.

Furthermore, for the emerging category of "virtual" crimes, decisions by online platforms, informed by algorithmic filtering, can shape the scope and meaning of the offenses they "host."[139] Thus, in the case of nonconsensual pornography, algorithmic preemption can come to dominate a significant portion of the information brought to the legal community's attention because it determines what content is subject to scrutiny and what materials are embedded into the endless sea of content available online.[140] To be sure, filtering need not be airtight to have a significant effect; that it prevents the accumulation of a critical mass of boundary-defying cases could be enough for it to have a profound adverse effect on the development of a vibrant normative debate.

A third line of criticism responds to the filter bubble theorists' assertion that filtering eliminates chance encounters with information deemed irrelevant.[141] Pariser describes this as lost encounters with information that becomes an "unknown unknown," meaning information that we do not even know that we are missing.[142] Accordingly, Sunstein frames his response to the filter bubble as a plea for serendipity in our encounters with new information.[143] Responding to concerns over the loss of chance encounters, Deven Desai avers that we are not really interested in serendipitous encounters with random unknown information because truly random information would be of little use.[144] What we do want, Desai argues, and what the proponents of the filter bubble should call for, is "better exposure to relevant, but

[139] On the effect of social media "social ordering," see Wu, *supra* note 7.

[140] On the idea of online obscurity, see Woodrow Hartzog & Evan Selinger, *Surveillance as Loss of Obscurity*, 72 WASH. & LEE L. REV. 1343 (2015).

[141] *See, e.g.*, Desai, *supra* note 124 (making this argument).

[142] *See* PARISER, *supra* note 9, at 106 ("In the filter bubble . . . . [y]ou don't see the things that don't interest you at all. You're not even latently aware that there are major events and ideas you're missing.").

[143] SUNSTEIN, #REPUBLIC, *supra* note 114, at 4–5.

[144] *See* Desai, *supra* note 124.

unknown information."[145] If this is indeed the case, what we really need is not less but *better* filtering, which could find for us the information we did not know fits our preferences.

Again, focus on personalized filtering hides the fact that this criticism holds true only if we believe the measure of relevance to be axiomatically valid.[146] Indeed, as long as we are interested only in information that best fits our personal preferences, we will be interested in serendipity only to the extent that chance encounters conform to our unrealized personal tastes. However, once we go beyond personal preferences, the measure of relevance itself can and sometimes should come into question, meaning that we also have an interest in encountering information that is genuinely irrelevant for our current purposes. As will be suggested in Part III, consent, for instance, may or may not be the right measure of relevance for the harms caused by unwanted distribution of sexual content. Filtering that uses consent as the measure of relevance will have the effect of hiding from sight information deemed immaterial to the finding of nonconsent, thus concealing the choice made between consent and other proxies for sexual autonomy. Ingenious filtering could, perhaps, shed new light on the meaning of consent, but it would not go as far as providing users with information deemed immaterial to the question of consent.

Finally, some assert that there is, in fact, nothing new about the emergence of filter bubbles, nothing that is not already apparent in the bounded nature of human rationality or the constricting effects inherent in social technologies such as law.[147] Still, although such constrictions, notoriously prevalent in legal analysis, far predate the use of computer algorithms, this criticism misses the point of the filter bubble theory. For this theory, a significant reason why filter bubbles are so potent is that they amplify existing human and bureaucratic failures and impede existing mechanisms of self-repair. Even when

[145] *Id.* at 560.

[146] *See* Gillespie, *supra* note 27, at 3 ("Then there are deeper problems with automating moderation, many of which resonate with familiar concerns about AI and data science more broadly, and that animate worries about automated policing, data-driven insurance assessments, hiring software, and automated medical diagnostics.").

[147] *Cf.* Jack M. Balkin, 2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data, *in* 78 OHIO ST. L.J. 1217, 1223 (2017) (arguing that algorithms are primarily inert media through which social relations between human beings take place); Gillespie, *supra* note 10, at 172 ("The particular patterns whereby information is either excluded from a database, or included and then managed in particular ways, are reminiscent of twentieth-century debates about the ways choices made by commercial media about who is systematically left out and what categories of speech simply don't qualify can shape the diversity and character of public discourse." (citation omitted)). I thank Jack Balkin, Josh Fairfield, and Daniel Markovits for highlighting this important argument.

algorithms introduce no new epistemological constraints, they can nonetheless act as "autopropaganda" mechanisms, exacerbating confirmation and selection biases as they transform these subconscious heuristics into design-based "non-choices" effectively hidden from the user.[148] Likewise, as the reliance on algorithmic filtering transforms those things users choose not to encounter into unknown unknowns, sparing them even the choice not to encounter them, filtering eliminates the existence of external pressures that could otherwise make users conscious of their choices and perhaps alter them.[149]

Similarly, bureaucratic path dependence can lead organizational epistemological structures to favor the familiar over the untried.[150] Introducing algorithmic filtering into path-dependent systems—and no system of thought seems more path dependent than law—can enhance this preexisting tendency to further pursue known concepts by making their pursuit more efficient, as the algorithm develops elaborate models that can scale any obstacles encountered down the road. At the same time, filtering also removes from sight any reminders that continuing down the familiar path is a choice not to take another. Filtering, so to speak, can hide from sight "the road not taken," and that can make all the difference.

B. *Algorithms and Legal Decision-Making*

Filter bubbles occur when algorithms control a bottleneck through which information passes to users. In personalized filtering, such bottlenecks are a direct consequence of the contemporary reliance on search engines, algorithmically curated news feeds, personalized social platforms, and the like. For legal decision-making, the control that algorithms have over the flow of information can result from their direct participation in legal proceedings[151] or from their influence on the normative universe from which legal decisions draw their information.[152] In both cases, an algorithm's decision about which

[148] *See, e.g.*, Gillespie, *supra* note 10, at 168–69, 177–79 (discussing algorithms' hidden epistemological effects); PARISER, *supra* note 9, at 29 (discussing algorithms' "autopropaganda" cognitive effects).

[149] *See* PARISER, *supra* note 9, at 84, 89–91 (discussing how algorithms hide conflicting information).

[150] On path dependence, see Hathaway, *supra* note 19; PARISER, *supra* note 9, at 133–34.

[151] *See generally* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7; RICHARDSON, SCHULTZ & SOUTHERLAND, *supra* note 6. On police use of algorithms, see Ferguson, *supra* note 7; Rich, *supra* note 7; Joh, *supra* note 6.

[152] *See* Bamberger, *supra* note 7 (discussing the systemic effects of the use of algorithms).

information is irrelevant to the normative domain can significantly affect law's development.

That legal filtering is at all conceivable is a consequence of the almost unfathomable strides machine-learning technology has taken in past decades, emulating one human skill after another, including intuitive and creative capabilities that until recently were thought to be impervious to algorithmic imitation.[153] To be sure, technological devices served legal ends long before machine learning. As Andrea Roth surveys, machines have long provided courts with information, including "opinions" conveyed by algorithmic expert systems.[154] Expert systems are essentially manually created algorithmic models that translate subject-matter expertise into formal-logic instructions executable by computer systems.[155] A familiar instance of such systems is tax software, in which professional understandings of tax laws and regulations are aggregated into a general model of tax accounting formed by a large number of formal if-then-else instructions incorporated into a user-friendly interface.[156]

In other cases, subject-matter expertise is used to manually devise the mathematical relations between various input variables and a desired dependent variable, relations modeled as a weighted mathematical formula.[157] A familiar, often notorious example for such algorithmic expert systems is the "risk scoring" algorithms used by police departments and courts for prioritizing investigations, making bail decisions, and sentencing.[158] The failings and harms of such systems have been extensively discussed in legal scholarship.[159] Of particular disrepute are the hidden biases plaguing these systems that are the results of processes similar to those discussed in Part I. Such systems

[153] The most striking recent example comes from GPT-3, a natural language–processing algorithm created by OpenAI, used to write entries for *The New York Times' Modern Love* section. *See* Cade Metz, *When A.I. Falls in Love*, N.Y. TIMES (Nov. 24, 2020), https://www.nytimes.com/2020/11/24/science/artificial-intelligence-gpt3-writing-love.html [https://perma.cc/8E5G-XXHS]. On the legal uses of machine learning, see Coglianese & Lehr, *supra* note 94, at 1161.

[154] Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 1981–82 (2017).

[155] On the history of expert systems, see NILSSON, *supra* note 31, at 229–50.

[156] *See* RUSSELL & NORVIG, *supra* note 29, at 22–24 (discussing expert systems).

[157] *Id.* at 676–86 (discussing linear regression).

[158] For discussions of predictive policing, see ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7, at 17; Simmons, *supra* note 48, at 1069–70. On risk scoring in bail decisions, see John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725 (2018). On sentencing, see Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017).

[159] *See, e.g.*, Citron & Pasquale, *supra* note 51; Ferguson, *supra* note 7; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018).

can also develop into unprecedented force multipliers, greatly expanding law enforcement agencies' reach and requiring appropriate regulatory tools to hold their users accountable for their newfound capabilities.[160]

Still, most expert systems involve legalistic filtering only in a derivative way.[161] Although the use of these systems can affect subsequent decision-making, their limited abilities do not allow them to determine what information is relevant to the decision, for all but the simplest forms of legal analysis involve nuanced decision-making models that are impervious to manual representation.[162] Modeling recidivism rates to predict risk is categorically different from modeling a police officer's decision as to whether a situation is "suspicious."[163] Traditional manually produced algorithms are simply incapable of doing the latter.

The advent of machine learning, however, is steadily overcoming this barrier, breaking new ground as it performs functions that are progressively closer to the heart of normative decision-making. This has been particularly true of narrowly defined, formal, and routine legal tasks.[164] Freed from the need to manually create and code decision-making models, advanced forms of machine learning can create dynamic and hyperdimensional decision-making models that mimic a totality-of-the-circumstances approach.[165] Such hyperdimensionality allows machine learning to produce holistic and nuanced models of legal concepts.[166] Deep methods of learning are capable of abstraction—they *can* see the forest for the trees.[167] Advanced case-based reasoning methods can operate in ways that are remarkably similar to legal

[160] *See* Joh, *supra* note 6, at 19; Simmons, *supra* note 48, at 1072.

[161] *See* Maggen, *supra* note 7 (distinguishing between fact-finding and legal-analysis algorithms).

[162] *See* Surden, *Artificial Intelligence*, *supra* note 6, at 1309 (discussing the limits of legal algorithms).

[163] *Cf.* Jason Millar & Ian Kerr, *Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots*, *in* ROBOT LAW 102 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2015) (discussing the limits of expert systems).

[164] *See* Pasquale, *supra* note 88, at 29; Surden, *Artificial Intelligence*, *supra* note 6, at 1309.

[165] *See* ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING 2 (3d ed. 2014); GOODFELLOW, BENGIO & COURVILLE, *supra* note 35, at 2–7 (discussing the use of "deep" models to emulate subjective decision-making).

[166] This does not mean that such a holistic model would be useful or even feasible. In fact, much effort is put into reducing the number of features the learning algorithm takes into account, in order to conserve computational efforts and avoid overfitting. *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 35, at 417; RUSSELL & NORVIG, *supra* note 29, at 751–54.

[167] *See* GOODFELLOW, BENGIO & COURVILLE, *supra* note 35, at 498.

analogy.[168] All these methods are used to produce algorithms that can reliably perform at least some aspects of legal analysis.[169]

To be sure, although the technology is constantly evolving, we are still a long way from algorithms that can engage in full-scale legal adjudication, primarily because of limitations imposed by insufficiently precise natural-language processing.[170] Still, the currently available algorithmic capabilities come in particularly handy as decision-makers struggle to filter the endless amounts of potentially relevant data the information age sends their way.[171]

The list of use cases in which machine learning assists law-related and other normative tasks is steadily growing. Lawyers increasingly apply machine learning in the course of preparing legal briefs, in so doing shaping subsequent legal proceedings.[172] Machine learning is used in electronic discovery proceedings, replacing human lawyers in sifting through documents in search of those relevant to a legal cause of action.[173] It aids legal research, filtering and categorizing relevant legal sources.[174] Federal agencies put machine learning to use to detect illicit behavior; an example is the SEC's employment of machine-learning algorithms to identify insider trading.[175] It is even used to vet legal strategies by evaluating their strength, meaning their relevance to the desired legal outcome.[176]

Perhaps the most extensive use of algorithmic filtering in legal matters is in online copyright adjudication, where massive amounts of user-uploaded content force platforms to heavily rely on the help of algorithms. As Maayan Perel and Niva Elkin-Koren illustrate, algorithms are commonly used in online copyright adjudications under the Digital Millennium Copyright Act as the first line of response to the

[168] *See* Janet L. Kolodner, *An Introduction to Case-Based Reasoning*, 6 A.I. REV. 3, 4 (1992).

[169] *See generally* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7.

[170] *See* Surden, *Artificial Intelligence*, *supra* note 6, at 1322–23.

[171] *See* Bamberger, *supra* note 7, at 707; Desai & Kroll, *supra* note 62, at 50–51. *See generally* R. GREGG DWYER ET AL., PROTECTING CHILDREN ONLINE: USING RESEARCH-BASED ALGORITHMS TO PRIORITIZE LAW ENFORCEMENT INTERNET INVESTIGATIONS, TECHNICAL REPORT (2016).

[172] *See* Dana Remus & Frank Levy, *Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law*, 30 GEO. J. LEGAL ETHICS 501 (2017).

[173] *See* Surden, *Artificial Intelligence*, *supra* note 6, at 1329–32.

[174] Some examples include CARA A.I., CASETEXT, https://casetext.com/ [https://perma.cc/42F6-C7WZ]; ROSS, *A Visual Guide to AI*, ROSS INTEL., https://www.rossintelligence.com [https://perma.cc/GK4V-2TUM]; SCHOLARSIFT, https://scholarsift.com [https://perma.cc/9PWQ-Y66A].

[175] *See* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7, at 22–25. *See generally* Ehret, *supra* note 7.

[176] *See* Surden, *Artificial Intelligence*, *supra* note 6, at 1331–32; Surden, *Machine Learning*, *supra* note 6, at 101–02.

vast number of automated notice and takedown requests platforms receive.[177] Similarly, platforms such as YouTube use algorithms to proactively detect uploaded content that infringes on copyrighted materials, allowing rights holders to object to the use or profit from it.[178] In this process, copyrighted materials are hashed—reduced to uniquely identifiable mathematical features—using a system called Content ID and matched, using a system called Copyright Match Tool, against any new upload to detect infringements.[179] Similar technology, named PhotoDNA, was developed in 2009 by Microsoft and Dartmouth College to help stop the spread of child pornography.[180] Like Content ID, PhotoDNA involves the hashing of images that have previously been marked as illegal to assist in locating copies and reproductions of these known images and preventing their continued distribution.[181]

As Wu suggests, although use of such assistive systems does not cede control over the decision to the algorithm, the algorithm's control over the initial stages of the process, either by proactively instigating it or by deciding which user complaints require human attention, creates hybrid adjudicative systems.[182] Such algorithmic gatekeeping, the filter bubble theory suggests, is bound to have a tacit effect not only on the subsequent decisions but also on decision-makers' states of mind and how they see the normative environment they operate in.

C. *The Rise of Legalistic Filters*

As the filter bubble theory suggests, the effect of algorithmic filtering is bound to the algorithm's method for determining the relevance of information.[183] Accordingly, it is important to notice that in most of the above examples, the algorithms make their determination in a manner that is more factual than normative;[184] therefore, the

[177] *See generally* Perel & Elkin-Koren, *supra* note 62; Perel & Elkin-Koren, *supra* note 7.

[178] *See, e.g.*, Wu, *supra* note 7, at 2007; *Copyright Management Tools*, YOUTUBE, https://support.google.com/youtube/topic/9282364?hl=en&ref_topic=2676339 [https://perma.cc/FPS2-6TK9].

[179] *How Content ID Works*, YOUTUBE, https://support.google.com/youtube/answer/2797370?hl=en&ref_topic=9282364 [https://perma.cc/AZW7-JTKN].

[180] *How Does PhotoDNA Technology Work?*, MICROSOFT, https://www.microsoft.com/en-us/photodna [https://perma.cc/U4G9-53QK].

[181] *Id.*

[182] *See* Wu, *supra* note 7, at 2008–20.

[183] *See generally* Gillespie, *supra* note 10.

[184] *See* Nicholas Thomas DeLisa, Note, *You(Tube), Me, and Content ID: Paving the Way for Compulsory Synchronization Licensing on User-Generated Content Platforms*, 81 BROOK. L. REV. 1275, 1281–82 (2016).

immediate concern with the effect of filtering revolves around potential inaccuracies and biases.[185] Undoubtedly, the line between fact-finding and legal classification is not all that clear; when an algorithm determines that new content is identical to material previously marked as copyrighted or prohibited, its measure of relevance is *similarity*, not legal classification, but it also has immediate effects on the meaning of the legal category involved.[186] Still, despite the considerable effect such quasi-normative measures of relevance can have, they are not *legalistic* in the sense discussed here, for the algorithm does not make its decision based on the material's *legal* relevance.

However, this reality is beginning to change, most vividly in the use of machine-learning algorithms to proactively detect objectionable content on social media. Until recently, social media content moderation relied heavily on user complaints, using algorithms mainly to assist human content moderators in dealing with enormous numbers of complaints.[187] However, in recent years, platforms have begun shifting toward fully algorithmic moderation, with the Covid-19 pandemic significantly accelerating this trend.[188] Today, these companies extensively rely on algorithms that *independently* determine whether content is potentially violative of the platforms' standards for prohibited content before it is subject to any human scrutiny. In March 2020, YouTube announced its implementation of new measures in which "automated systems will start removing some content without human review," detecting "potentially harmful content and then send[ing] it to human reviewers for assessment."[189] Likewise, in April of the same year, Twitter began using algorithms trained on moderation decisions to "surfac[e] content that's most likely to cause harm and should be reviewed first" and "proactively identify rule-breaking

[185] *See* Katrina Geddes, *Meet Your New Overlords: How Digital Platforms Develop and Sustain Technofeudalism*, 43 COLUM. J.L. & ARTS 455, 461–65 (2020).

[186] *See, e.g.*, Amy Kapczynski, *The Cost of Price: Why and How to Get Beyond Intellectual Property Internalism*, 59 UCLA L. REV. 970 (2012); *cf.* LANGDON WINNER, *Do Artifacts Have Politics?*, *in* THE WHALE AND THE REACTOR: A SEARCH FOR LIMITS IN AN AGE OF HIGH TECHNOLOGY 19 (1986).

[187] *See, e.g.*, Desai & Kroll, *supra* note 62, at 51.

[188] *See* Elizabeth Dwoskin & Nitasha Tiku, *Facebook Sent Home Thousands of Human Moderators Due to the Coronavirus. Now the Algorithms Are in Charge*, WASH. POST (Mar. 24, 2020), https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus [https://perma.cc/DL88-SX9V]; James Vincent, *Facebook Is Now Using AI to Sort Content for Quicker Moderation*, VERGE (Nov. 13, 2020, 9:00 AM), https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation [https://perma.cc/JUS5-UMKW].

[189] *Protecting Our Extended Workforce and the Community*, YOUTUBE BLOG (Mar. 16, 2020), https://blog.youtube/news-and-events/protecting-our-extended-workforce-and [https://perma.cc/BQ8A-K4AM].

content before it's reported."[190] Similarly, Facebook has steadily increased its reliance on proactive filtering used to identify materials that infringe on its community standards before they are reported.[191] In 2021, over ninety-five percent of all hate speech violations on Facebook were proactively detected, with algorithms independently determining what speech falls under this classification.[192] Finally, in the first quarter of 2021, YouTube reported using automated flagging to remove about nine million videos, with fewer than half a million removals originating from human sources.[193]

Beyond social media, the use of algorithms that draw on legal categories has been prevalent in the prevention of child pornography. In 2018, Google announced its development of an algorithm, made freely available in the form of an API titled "Content Safety," capable of *originally* identifying materials falling under the category of child pornography.[194] Google presented the Content Safety API as a screening tool to be used prior to any human evaluation of the materials, with the purpose of minimizing human contact with disturbing materials and scaling up human adjudication.[195] Although Google has not disclosed information on how its algorithm detects child sexual abuse, it suggests that it does so through the use of machine-learning classifiers.[196] In 2021, Pornhub, responding to mounting public pressure in response to a 2020 *New York Times* piece exposing its facilitation of illegal and

[190] Vijaya Gadde & Matt Derella, *An Update on Our Continuity Strategy During COVID-19*, TWITTER BLOG (Apr. 1, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19 [https://perma.cc/XBY6-Q3QF].

[191] *How Facebook Uses Super-Efficient AI Models to Detect Hate Speech*, META AI (Nov. 19, 2020), https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech [https://perma.cc/8WHJ-JKRE].

[192] *Hate Speech*, META TRANSPARENCY CTR., https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook (last visited Apr. 22, 2022).

[193] *YouTube Community Guidelines Enforcement*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2021Q1;exclude_automated:human_only&lu=total_removed_videos (last visited Apr. 22, 2022).

[194] Nikola Todorovic & Abhi Chaudhuri, *Using AI to Help Organizations Detect and Report Child Sexual Abuse Material Online*, GOOGLE BLOG (Sept. 3, 2018), https://www.blog.google/around-the-globe/google-europe/using-ai-help-organizations-detect-and-report-child-sexual-abuse-material-online [https://perma.cc/K2SR-7JGW].

[195] *Fighting Child Sexual Abuse Online*, GOOGLE, https://static.googleusercontent.com/media/protectingchildren.google/en//static/pdf/content-safety-api.pdf [https://perma.cc/8HZR-EMJN].

[196] Kristie Canegallo, *Our Efforts to Fight Child Sexual Abuse Online*, GOOGLE BLOG (Feb. 24, 2021), https://blog.google/technology/safety-security/our-efforts-fight-child-sexual-abuse-online [https://perma.cc/36D8-HAWL].

exploitative materials,[197] announced its adoption of "industry-leading measures for verification, moderation and detection," which would be implemented across the properties of its parent company, MindGeek, which controls a significant portion of the online pornography production market.[198] These measures, the pornography colossus announced, will include proactive screening involving manual human review and "a variety of automated detection technologies," including Google's Content Safety API.[199]

Admittedly, the algorithmic legal analysis that goes into the detection of child pornography can be minimal, as the very appearance of a child in a video containing sexual imagery is a strong indication of illegality. The same, however, cannot be said of Facebook's 2019 introduction of algorithmic filtering to determine whether the uploading of sexual material is consensual. Like other platforms, Facebook, in making this determination, initially relied on complaints and human moderation and used algorithms mainly to take down materials that were already marked as nonconsensual.[200] In 2019, however, it began using an algorithm trained on past takedown decisions to independently develop a model of nonconsent and using this model to detect nonconsensual distribution before it is seen by anyone.[201] As Antigone Davis, Global Head of Safety at Facebook, revealed, Facebook seeks to expand the use of this technology in collaboration with other companies, such as Twitter, YouTube, Microsoft, Snap, and Reddit.[202] Such cooperation, Davis implies, would be modeled after these companies' cooperation, in relation to similar technologies used to detect and prevent terrorist propaganda, in announcing their intention to "exchange best practices" as they "develop and implement new content detection and classification techniques using machine learning."[203] There is, therefore, good reason to believe that the algorithmic preemptive prevention of nonconsensual

[197] Nicholas Kristof, *An Uplifting Update, on the Terrible World of Pornhub*, N.Y. TIMES (Dec. 9, 2020), https://www.nytimes.com/2020/12/09/opinion/pornhub-news-child-abuse.html [https://perma.cc/7SAS-NP7R].

[198] *Pornhub Sets Standard for Safety and Security Policies Across Tech and Social Media; Announces Industry-Leading Measures for Verification, Moderation and Detection*, PORNHUB (Feb. 2, 2021), https://www.pornhub.com/press/show?id=2172 [https://perma.cc/ERE4-722Z].

[199] *Id.*

[200] *See infra* Section III.C.

[201] *See* Davis, *supra* note 2.

[202] *See* Solon, *supra* note 36.

[203] *Global Internet Forum to Counter Terrorism*, TWITTER BLOG (June 26, 2017), https://blog.twitter.com/official/en_us/topics/company/2017/Global-Internet-Forum-to-Counter-Terrorism.html [https://perma.cc/NJA9-RQJ4]; *see also* Solon, *supra* note 36.

distribution would become an industry best practice, taking hold over all mainstream platforms.[204]

In many ways, Facebook's algorithm is the clearest example to date of a filtering algorithm prone to creating a legalistic filter bubble. By independently determining the legal meaning of the screened information, this filtering sets the boundaries for consequent human decisions. Furthermore, much more so than any other form of content moderation, this initial flagging can lead to criminal charges and civil suits being brought against the uploader, thus setting the tone for subsequent legal proceedings. Since nonconsensual distribution of sexual materials primarily occurs through online intermediaries, the very meaning of this offense will be determined by the decisions made by the algorithms in these cases.

Although this degree of reliance on legalistic filtering is not yet dominant directly in legal proceedings, there is reason to believe that it is only a matter of time before algorithmic filtering expands from the virtual domain to legal decision-making, especially in routine adjudications where unassisted decision-making can result in intolerable backlogs.[205] As I discuss elsewhere, algorithmic systems are currently employed by child protective services agencies, with the intention of employing machine-learning analysis to help reduce their unbearable workloads, to triage complaints of child maltreatment.[206] Similarly, a collaboration between Stanford's Regulation, Evaluation, and Governance Lab and Carnegie Mellon's Language Technologies Institute is currently developing an algorithmic decision support system meant to assist the Board of Veterans Appeals in its mass adjudication of disability or veterans' benefits determinations.[207] In the most striking example thus far, the Brazilian judiciary is in the process of implementing machine-learning triaging systems to assist in addressing the country's immense judicial backlog.[208]

In current and emerging use cases, the filter bubble theory suggests that reliance on algorithms that use legal categories to make the filtering decisions will shape the worldviews of those reliant on it in accordance

[204] *See* Bamberger, *supra* note 7, at 712–13 (discussing institutional isomorphism).

[205] *See* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7, at 10 (examining the use of algorithms by federal agencies).

[206] *See* Maggen, *supra* note 7.

[207] Daniel E. Ho & Matthias Grabmair, *Toward a Decision Support System for Veterans Adjudication*, STANFORD L. SCH., https://law.stanford.edu/event/codex-speaker-series-dan-ho [https://perma.cc/9MAU-D6BV].

[208] *See* KATIE BREHM ET AL., NAT'L COUNCIL OF JUST. INST. FOR TECH. & SOC'Y OF RIO DE JANEIRO, THE FUTURE OF AI IN THE BRAZILIAN JUDICIAL SYSTEM: AI MAPPING, INTEGRATION, AND GOVERNANCE (2020), https://www.sipa.columbia.edu/academics/capstone-projects/ai-driven-innovations-brazilian-judiciary [https://perma.cc/E6NL-HKX2].

with the legalistic measure of relevance. This effect can vary according to the specifics of the filtering mechanism, the precise function that animates it, and how this function is attained from the training sets. Still, as I argue in the following pages, legalistic filtering has as a main feature baking in the dominant legal paradigms that inform its measure of relevance and obscures anything that falls outside of them.

D. *A Holmesian Filter*

Filter bubbles produce their effects by reinforcing the user's acceptance of the measure of relevance and reducing encounters with information that undermines it. In personalized filtering, filter bubbles ensure that users are given only information deemed relevant to their personal tastes and interests and are cut off from opposing views; the societal harms of such bubbles are arguably the radicalization and polarization this constriction produces. In the case of legalistic filtering, limiting decision-makers' vantage points to legally relevant information and "invisibly hiding" legally immaterial information is prone to engendering normative ossification, entrenching dominant legal paradigms, and suppressing debate on their appropriateness and decency.[209]

The constricting of decision-makers' worldviews is akin to making them into Holmesian "bad men." Oliver Wendell Holmes famously put forward the bad man's view of law to emphasize the importance of adopting an amoral, reductive view of legal meaning akin to the viewpoint of the proverbial "bad man" who views legal rules through the single prism of the likelihood of facing official sanction.[210] Legal reasoning, Holmes sought to remind us, is inherently recursive, and it is folly to assign it any moral or otherwise extralegal considerations. As if channeling this view, legal filter bubbles limit decision-makers' normative world to legally relevant information, stripping it of anything irreducible to its legal bottom line.

Holmes did not, however, intend to suggest that this reductionist legalism exhausts the normative space that Law occupies. Rather, like other ardent positivists,[211] Holmes accentuated law's amorality to underscore the need for legal adjudicators to supplant positive law with extralegal considerations drawn from a social-scientific appreciation of

[209] *Cf.* Gillespie, *supra* note 27, at 3–4; Green & Chen, *supra* note 48, at 4.

[210] Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 459–61 (1897).

[211] *See, e.g.*, SCOTT J. SHAPIRO, LEGALITY 99–102 (2011).

the social reality in which the law operates.[212] The fact that legal reasoning is inherently limited to positive law is precisely why legal adjudication must constantly look outside it, when determining law's content and development, to the social advantages legal norms are meant to produce. For Holmes, the duty to take these extralegal considerations into account—the duty to transform law into Law, so to speak—rests with the courts. The exercise of this duty, Holmes believed, is an "inevitable" part of legal adjudication, so "the result of the often proclaimed judicial aversion to deal with such considerations is simply to leave the very ground and foundation of judgments inarticulate, and often unconscious."[213]

As Holmes keenly noted, for adjudicators to actively exercise their duty to go beyond law, they must first be made aware of law's outer limits, lest they passively leave the social considerations that shape its course untouched.[214] Legal filter bubbles not only hide law's outer limits, they do so invisibly, desensitizing adjudicators to the inert regressiveness of their decisions. The better legal filtering algorithms become at emulating strictly legalistic decision-making, the more likely they are to have this effect on human decision-makers as adjudicators increasingly rely on filtered information and remain oblivious to the existence of cases that evade the grasp of prevalent norms. To this effect, legalistic filter bubbles not only hinder adjudicators' ability to consciously decide law's path by subjecting existing legal paradigms to external scrutiny but also strip the rich legal substance of past decisions of their social meaning, reducing it to barren legal models.[215] In comes Law, out goes law.

If, therefore, the fault lies with filtering algorithms that measure relevance according to an impoverished, legalistic measure of relevance, could the answer to this problem be a turn to a broader notion of Law, one that incorporates greater parts of the social reality in which it operates?

The problem with this solution is that, as Section I.A described, all too often, the only readily available source of training data for the creation of law-related algorithms is past decisions.[216] In transforming past decisions into models of legal concepts, machine learning

[212] *See* E. Donald Elliott, *Holmes and Evolution: Legal Process as Artificial Intelligence*, 13 J. LEGAL STUD. 113, 115 (1984).

[213] *See* Holmes, *supra* note 210, at 467.

[214] *Cf.* SHAPIRO, *supra* note 211, at 398–400.

[215] *See* Davis, *supra* note 62, at 189.

[216] *See* Steven A. Israel, Philip Sallee, Franklin Tanner, Jonathan Goldstein & Shane Zabel, *Applied Machine Learning Strategies*, 39 IEEE POTENTIALS 38, 38 (2020); ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 7, at 19.

essentially embraces the Holmesian shift from logic to experience. As mentioned, before the advent of machine learning, modeling human behavior was an exercise in formal logic, as programmers were required to transform subject-matter expertise into clearly defined rules. As if taking its cues from Holmes, machine-learning modeling broke with the path of logic and took inference from experience to be its animating principle, resulting in two significant effects. First, using machine learning to extract legal concepts from past decisions requires the datafication of these decisions.[217] This entails reducing rich legal texts planted in a social reality into indexes of variables judged only on their connection to the output variable. The outward expression recorded in the modeled datasets is already only a derivative manifestation of legal adjudication; modeling it can at best produce an impoverished secondhand model of legal reasoning.[218] Frank Pasquale and Glyn Cashwell refer to this reduction as the creation of a "jurisprudence of behaviorism" that overemphasizes the importance of measurable input data, thus distorting the ensuing model.[219]

Second, and more importantly, using *supervised* learning to extract legal meaning brings algorithmic modeling even closer to Holmesian legalistic analysis. Ascertaining the meaning of law, Holmes advised, is synonymous with predicting how legal decision-makers would rule in a given case in light of past decision-making patterns; this is precisely how supervised learning models the meaning of classifications: by connecting patterns in past decisions to the applicable label.[220] Supervised machine learning thus inevitably reduces the meaning of any piece of information to its connection to a single, unequivocal legal classification.[221] By doing so, algorithmic modeling essentially turns the labels assigned in past decisions into immovable Archimedean points grounding the algorithm's operation; hence, any attempt to broaden the algorithm's measure of relevance requires training it on richer data with labels correlating to more profound notions of human flourishing.

[217] On datafication, see MAYER-SCHÖENBERG & CUKIER, *supra* note 32, at 73–97; KELLEHER & TIERNEY, *supra* note 33, at 46.

[218] As Marshall McLuhan suggests, it may be true of any technology that it takes as input the products of another technology. *See* MCLUHAN, *supra* note 123, at 8, 56; *see also* Kroll et al., *supra* note 6, at 646; Pasquale, *supra* note 88, at 3.

[219] *See* Pasquale & Cashwell, *supra* note 62, at 65.

[220] *See* Holmes, *supra* note 210; *see also* Surden, *Artificial Intelligence*, *supra* note 6, at 1331.

[221] *See* SIMONE BROWNE, DARK MATTERS: ON THE SURVEILLANCE OF BLACKNESS 114 (2015).

Unfortunately, if such data exists at all, it is certainly not available in sufficient quantities.[222]

As already noted, given machine learning's insatiable appetite for massive amounts of data, designers are routinely forced to accept poor or limited proxies for the algorithm's actual purpose when no better source is available.[223] Likewise, for most law-related algorithms created through supervised learning, the unavailability of adequate training data other than past decisions precludes the development of filtering algorithms not based on a legalistic measure of relevance. As the filter bubble theory suggests, the result of such filtering is the entrenchment of prevailing legal concepts and the suppression of normative debate. As we shall now see, this is precisely what might occur with the fight against nonconsensual pornography.

III. FILTERING NONCONSENSUAL DISTRIBUTION

The harmful distribution of sexual materials without the consent of those depicted in them is certainly not new. However, the technologies of the information age have given it unique urgency as well as recognition as a grave violation of the victim's sexual autonomy. Still, despite this rapid legal acknowledgment, or perhaps because of it, what harms the prohibition of nonconsensual distribution aims to prevent are not entirely obvious: the victim's proprietary interests in the images, their reputational interests, their privacy, emotional well-being, or sexual autonomy, or all of the above. In this nascent state, the discussion of illicit distribution is comparable to the state of the discussion on sexual assault when it revolved around stranger rape, the clearest but also least common form of sexual assault.[224] As was the case with its real-world counterpart, the discussion of virtual violations of sexual autonomy is currently almost exclusively focused on the extreme cases, where victims did not know of the perpetrator's intention to distribute the materials and did not acquiesce to it in any way. However, unlike the burgeoning discussion that today surrounds the meaning of real-world sexual violations, the emergence of legalistic filter bubbles

[222] In different settings, designers commonly rely on organically created labeled data, such as tagged pictures on social media. *See* KELLEHER, *supra* note 30, at 21–22. On manufactured databases painstakingly labeled, often employing low-paid human labor, see Kate Crawford & Vladan Joler, *Anatomy of an AI System*, SHARE LAB & AI NOW INST. (Sept. 7, 2018), https://anatomyof.ai [https://perma.cc/2RBP-EGZ4].

[223] *See* Elkin-Koren, *supra* note 7, at 5.

[224] *See generally* SUSAN ESTRICH, REAL RAPE (1987).

threatens to preserve the discussion of virtual violations in its embryonic state.

A. *The Many Forms of Violative Distribution*

The wrong involved in the violative distribution of sexual images has taken different forms with different names that reflect varying legal conceptions, notions of harm, and degrees of wrongdoing. The development of these forms has often been a consequence of technological developments. In one of its earliest modern manifestations, the printing press was used to mass-produce pamphlets weaponizing sexuality to strike at influential female figures, such as the revolutionary distribution of sexual depictions of Marie Antoinette.[225] Fast forward to the end of the nineteenth century, and new printing technologies and the development of portable camera equipment created a new form of unrelenting journalism, famously leading Samuel Warren and Louis Brandeis to put forward new conceptions of privacy to account for the harms wrought by eager journalists equipped with handheld cameras.[226] When, decades later, photographers have captured on film moments in which individuals are accidentally exposed in public in so-called wardrobe malfunctions and media outlets have published these photos, courts have addressed such incidents as violations of privacy, following the path set by the two.[227] However, although courts recognized the harm such publications did to the victims' privacy interests, they generally absolved newspapers of liability when the publication was deemed newsworthy.[228]

Technological advancements were also behind the notorious cases in which the pornographic magazine *Hustler* published images of naked women against their will. In the 1980s, the magazine published a section encouraging women to send it their naked images for publication, relying on technological developments such as instant polaroid cameras and automated film development that made it easier for

[225] *See* Lynn Hunt, *The Many Bodies of Marie Antoinette: Political Pornography and the Problem of the Feminine in the French Revolution*, *in* EROTICISM AND THE BODY POLITIC 108, 110 (Lynn Hunt ed., 1991).

[226] Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193 (1890).

[227] S*ee, e.g.*, Daily Times Democrat v. Graham, 162 So. 2d 474 (Ala. 1964) (finding a newspaper liable for intrusion of privacy for publishing images of a woman involuntarily exposed in a county fair).

[228] *See, e.g.*, McNamara v. Freedom Newspapers, Inc., 802 S.W.2d 901 (Tex. Ct. App. 1991) (ruling that the newspaper is immune from liability.)

nonprofessionals to capture and develop intimate images.[229] When a couple's photographs were copied by the developer and a woman's Polaroids were stolen, and both ended up being published by *Hustler*, courts found that the magazine was negligent in its efforts to ensure the consent of those depicted.[230] It is, however, noteworthy that the courts found that the magazine was liable not because its publication of their intimate images harmed the women but rather because the publication created the false impression that the women *consented* to the publication of their images in the pornographic magazine.[231]

The 1990s saw the move from analog to digital video equipment, making it easier to both create home videos discreetly and to mass-reproduce them. Together with the advent of the internet, the turn of the century was the age of "celebrity sex tapes," with home videos of public figures broadly distributed against their will.[232] When such cases reached courts, adjudication often revolved around the commercialization of the distribution and its appropriation of the victims' copyrights and right of publicity.[233] Often, these abuses ended in settlements that transformed the unlawfully distributed materials into consensual pornography.[234]

Continuing this trend, the twenty-first century brought with it Web 2.0 as websites gravitated toward user-created content. At the same time, ever-shrinking cameras made it easier to surreptitiously capture sexual images of people unaware they were being filmed and anonymously distribute the images online. To fight this phenomenon, Congress passed the Video Voyeurism Prevention Act of 2004, which made it a misdemeanor to intentionally capture a person's sexual images without their consent when they have a reasonable expectation of privacy.[235]

Finally, the arrival of smartphones put the ability to create pictures and videos at almost every person's fingertips, making the consumption and sharing of sexual content a salient feature of contemporary

[229] *See* Ashby v. Hustler Mag., Inc., 802 F.2d 856 (6th Cir. 1986) (regarding Polaroids); Wood v. Hustler Mag., Inc., 736 F.2d 1084 (5th Cir. 1984) (discussing automatic development).

[230] *Ashby*, 802 F.2d 856; *Wood*, 736 F.2d 1084.

[231] *Ashby*, 802 F.2d at 858 (discussing Hustler's liability for falsely representing that a woman gave consent for the publication of her sexual images in the pornographic magazine); *Wood*, 736 F.2d at 1089–90 (discussing Hustler's false light liability).

[232] *See* Lola Ogunnaike, *Sex, Lawsuits and Celebrities Caught on Tape*, N.Y. TIMES (Mar. 19, 2006), https://www.nytimes.com/2006/03/19/fashion/sundaystyles/sex-lawsuits-and-celebrities-caught-on-tape.html [https://perma.cc/ZBU9-YUVU].

[233] *See, e.g.*, Michaels v. Internet Ent. Grp., 5 F. Supp. 2d 823 (C.D. Cal. 1998) (granting preliminary injunction in favor of plaintiff).

[234] *See* Ogunnaike, *supra* note 232.

[235] 18 U.S.C. § 1801.

sexuality. With only a few clicks separating real-world from online sexuality, the virtual domain became a microcosm (or macrocosm) of interpersonal sexuality, replicating its potential for both self-expression and abuse.[236] Online venues soon became hotbeds for human trafficking.[237] Similarly, "sextortion" became a term to describe the migration of the ancient practice of extorting sexual acts into the virtual domain, with perpetrators forcing victims to provide them with sexual images, which they then leveraged to extort more images and often more violative sexual content by threatening to otherwise distribute the victim's sexual images—and at times making good on these threats.[238]

These technological developments also created the most recent form of violative distribution. In the early 2000s, websites hosting user-uploaded pornographic content, as well as mainstream social media platforms, saw a rise in sexual videos and images uploaded by former sexual partners of the persons depicted without the latter's knowledge or agreement—a phenomenon that became known as "revenge porn."[239] Not long after that, websites with the explicit or implicit intention of profiting off such content began popping up and using it to attract user traffic and charge victims sizable fees to take down their images.[240] Such malicious sites, however, have not remained the sole source of distribution, as perpetrators often use social media platforms to specifically target the victim's acquaintances.[241] At other times, social platforms are used as the means for exchanging images between perpetrators and others without the explicit intention of reaching or affecting their unwitting victims.[242] As the phenomenon grew in scope and the extent of the destruction it causes its victims became apparent, it gained greater scholarly and legal attention and began to be

[236] *See* I. India Thusi, *Reality Porn*, 96 N.Y.U. L. REV. 738 (2021).

[237] *See* Melissa Farley, Kenneth Franzblau & M. Alexis Kennedy, *Online Prostitution and Trafficking*, 77 ALB. L. REV. 1039 (2013).

[238] *See* BENJAMIN WITTES, CODY POPLIN, QUINTA JURECIC & CLARA SPERA, BROOKINGS, SEXTORTION: CYBERSECURITY, TEENAGERS, AND REMOTE SEXUAL ASSAULT (2016), https://www.brookings.edu/wp-content/uploads/2016/05/sextortion1-1.pdf [https://perma.cc/SJH8-XV9W].

[239] *See* Alexa Tsoulis-Reay, *A Brief History of Revenge Porn*, N.Y. MAG. (July 19, 2013), https://nymag.com/news/features/sex/revenge-porn-2013-7 (last visited Apr. 23, 2022).

[240] *See* Mary Anne Franks, *"Revenge Porn" Reform: A View from the Front Lines*, 69 FLA. L. REV. 1251, 1255, 1272 n.148 (2017).

[241] *See* Roni Rosenberg & Hadar Dancig-Rosenberg, *Reconceptualizing Revenge Porn*, 63 ARIZ. L. REV. 199, 205–06 (2021).

[242] *See id.*

commonly referred to as "nonconsensual pornography."[243] Although others suggest the more appropriate term is "image-based sexual abuse,"[244] I will generally use the former term as it currently dominates legal and scholarly discourse.

B. *The Legal Response to Nonconsensual Pornography*

The legal response to nonconsensual pornography has been split between reliance on civil law instruments to compensate victims for the harms they suffered and force platforms to take down the images, and the use of criminal prohibitions to target malicious websites and perpetrators and deter would-be distributors. As a civil law cause of action, nonconsensual distribution is conceived of as a violation of the victim's right to privacy. In the past, when the victims of unwanted distribution were mainly celebrities, and the motives for dissemination were monetary, legal proceedings gravitated toward compensating the victims for the infringement of their proprietary rights in their public image and their intellectual property as the material's creators.[245] With victims now being largely nonpublic figures, the dominant cause for action is an invasion of privacy as delineated by the *Restatement (Second) of Torts*, with the distribution of nonconsensual sexual images considered to be "highly offensive to a reasonable person" and "not of legitimate concern" to the public.[246]

Still, in a considerable number of cases, the privacy cause of action can be questioned by distributors when victims voluntarily share their explicit images with them or agree to have their images captured.[247] In response to this line of defense, courts and scholars underscore the contextual nature of consent, stressing that by willingly relinquishing part of their privacy as they voluntarily share their images, victims do not forfeit their right to not have their images publicly distributed.[248] As Danielle Citron and Mary Anne Franks suggest, victims' consent to grant someone possession of their sexual images does not imply their consent for others to also see the images, and the violation of privacy

[243] *See* Franks, *supra* note 240, at 1258; Ari Ezra Waldman, *A Breach of Trust: Fighting Nonconsensual Pornography*, 102 IOWA L. REV. 709 (2017); Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 346 (2014); Emily Poole, Comment, *Fighting Back Against Non-Consensual Pornography*, 49 U.S.F. L. REV. 181 (2015).

[244] *See* McGlynn & Rackley, *supra* note 22.

[245] *See, e.g.*, Michaels v. Internet Ent. Grp., 5 F. Supp. 2d 823 (C.D. Cal. 1998).

[246] RESTATEMENT (SECOND) OF TORTS § 652D (AM. L. INST. 1977).

[247] *See, e.g.*, Pohle v. Cheatham, 724 N.E.2d 655, 658–59 (Ind. Ct. App. 2000).

[248] *See id.* at 661; Citron & Franks, *supra* note 243, at 348.

revolves around this latter meaning of consent.[249] For others, such as Ari Ezra Waldman, the difficulty U.S. privacy law has in responding to this contextual nature of consent invites the adoption of the breach-of-trust privacy tort, prevalent in the United Kingdom, which more closely captures the gist of the perpetrator's wrongdoing.[250]

Others suggest turning to copyright law to assist victims in forcing platforms to take down their images. Platforms have little legal incentive to promptly respond to takedown requests, as they are immunized from liability to harms caused by content they host as a result of § 230 of the Communications Decency Act (CDA).[251] The CDA was enacted partly in response to the New York Supreme Court's ruling in *Stratton Oakmont v. Prodigy Services Co.*,[252] in which the court held that by removing materials it deemed offensive and in "bad taste" from its service, Prodigy "arrogated to itself the role of determining what is proper for its members to post and read," making it liable as a publisher to the harms caused by the materials it hosts. Section 230 sought to encourage such "Good Samaritan" content moderation by immunizing internet service providers from consequent publisher liability, but it was eventually interpreted as shielding websites from almost all civil liability—excepting infringements of intellectual property law.[253] Accordingly, when victims are those who created the images, they have a protected proprietary interest in the images that is unaffected by § 230; consequently, authors such as Amanda Levendowski suggest turning to the instruments put into use by the Digital Millennium Copyright Act to force platforms to punctually remove infringing content.[254]

In addition to civil remedies, states have taken an almost unanimously resolute stand in using criminal law to condemn at least some form of nonconsensual distribution.[255] Initially, the main targets of criminal proceedings were the operators of malicious "revenge porn" websites, who were charged with committing crimes incidental to the distribution.[256] Over time, states began criminalizing nonconsensual

[249] Citron & Franks, *supra* note 243, at 348.

[250] Waldman, *supra* note 243.

[251] 47 U.S.C. § 230.

[252] Stratton Oakmont, Inc., v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

[253] *See* 47 U.S.C. § 230(c); Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401 (2017).

[254] Amanda Levendowski, Note, *Using Copyright to Combat Revenge Porn*, 3 N.Y.U. J. INTELL. PROP. & ENT. L. 422 (2014).

[255] *See* Rosenberg & Dancig-Rosenberg, *supra* note 241, at 202.

[256] *See* Phil Helsel, *Revenge Porn Kingpin Hunter Moore Pleads Guilty, Faces Jail*, NBC NEWS (Feb. 25, 2016, 11:03 PM), https://www.nbcnews.com/news/crime-courts/revenge-porn-

distribution itself, with forty-six states and the District of Columbia currently explicitly proscribing at least some elements of it.[257] One of the significant points of divergence between these statutes is their scienter requirements, with some states requiring proof that the perpetrator intended to harass or cause the victim emotional harm and others making the offender's awareness of the distribution's nonconsensuality the pertinent point.[258]

Judicial discussions of these prohibitions' constitutionality are likewise split between courts that view nonconsent as this wrong's gravamen and those that focus on the wrongdoer's malicious intent. Both approaches have thus far refused to engage with the complex meaning of consent. Often the question courts address is whether the prohibition's curtailment of free speech is narrowly tailored to the state's interest in preventing the harm of unwanted distribution. The Wisconsin Court of Appeals in *State v. Culver*[259] upheld the state's prohibition despite the absence of a malicious intent requirement, reasoning that the nonconsensuality requirement sufficiently limits the prohibition's reach and that adding a malicious intent requirement would add little to better tailor the prohibition to the harm it is meant to address.[260] The same reasoning informed the Illinois Supreme Court's decision in *People v. Austin*,[261] in which it held that the Illinois statute "implicitly includes an illicit motive or malicious purpose" and is therefore sufficiently constrained.[262] The *Austin* court, however, refused to engage with questions that complicate the meaning of consent, ruling that they are to be answered on a case-by-case basis.[263] A similarly simplistic approach to consent was adopted by the Supreme Court of Minnesota in upholding the state's prohibition; the court wrote that "[i]n our view, it is not difficult to obtain consent before disseminating a private sexual image. Simply ask permission."[264]

kingpin-hunter-moore-pleads-guilty-faces-jail-n313061 [https://perma.cc/7PDF-BYFT]; Andrea Peterson, *Alleged Revenge Porn Web Site Operator Arrested in California*, WASH. POST (Dec. 10, 2013), https://www.washingtonpost.com/news/the-switch/wp/2013/12/10/alleged-revenge-porn-web-site-operator-arrested-in-california [https://perma.cc/LP2V-H4LG].

[257] *See 46 States + DC + One Territory Now Have Revenge Porn Laws*, CYBER C.R. INITIATIVE, https://www.cybercivilrights.org/revenge-porn-laws [https://perma.cc/V48A-SJN4].

[258] *See* Katherine G. Foley, Note, *"But, I Didn't Mean to Hurt You": Why the First Amendment Does Not Require Intent-to-Harm Provisions in Criminal "Revenge Porn" Laws*, 62 B.C. L. REV. 1365 (2021).

[259] State v. Culver, 918 N.W.2d 103 (Wis. Ct. App. 2018).

[260] *Id.* at 111.

[261] People v. Austin, 155 N.E.3d 439, 471 (Ill. 2019), *cert. denied*, 141 U.S. 233 (2020).

[262] *Austin*, 155 N.E.3d at 471 (citing *Culver*, 918 N.W.2d 103).

[263] *Id.* at 469.

[264] State v. Casillas, 952 N.W.2d 629, 644 n.9 (Minn. 2020).

However, other than in extreme cases, consent is hardly a clear, straightforward term, as the copious scholarship on its meaning for real-world sexual violations can attest.[265]

In contrast, the Texas Court of Appeals held that Texas's prohibition was unconstitutional because "its application is not attenuated by the fact that the disclosing person had no intent to harm the depicted person or may have been unaware of the depicted person's identity."[266] The Texas legislature amended its prohibition in response.[267] Likewise, the Vermont Supreme Court emphasized in its decision to uphold the state's prohibition the statute's inclusion of a "rigorous intent element" requiring "a specific intent to harm, harass, intimidate, threaten, or coerce the person depicted or to profit financially."[268]

In these latter opinions, the courts seem to follow in the footsteps of past decisions regarding real-world sexual abuse, in which courts viewed the use or threat of physical force as the distinguishing mark of prohibited violations.[269] Worried about a slippery slope and unwilling to confront the thorny question of sexual consent, courts have often relied on physical force as a bright-line distinction between unlawful coercion and the myriad ways in which people can bring others to acquiesce to unwanted sexual contact.[270] Similarly, even though courts and legislatures today are undoubtedly aware of the terrible harms caused by unwanted distribution regardless of the distributor's intent, many courts and legislatures seem reluctant to engage with the challenges raised by the consent paradigm, instead singling out those clear cases in which the perpetrator intended to harm the victim.

[265] For a number of scholarly attempts to engage with the complex meaning of consent to sexual contact, see ALAN WERTHEIMER, CONSENT TO SEXUAL RELATIONS (Gerald Postema ed., 2003); PETER WESTEN, THE LOGIC OF CONSENT: THE DIVERSITY AND DECEPTIVENESS OF CONSENT AS A DEFENSE TO CRIMINAL CONDUCT (2004); Donald A. Dripps, *Beyond Rape: An Essay on the Difference Between the Presence of Force and the Absence of Consent*, 92 COLUM. L. REV. 1780 (1992); Mark Dsouza, *Undermining Prima Facie Consent in the Criminal Law*, 33 LAW & PHIL. 489 (2014); William N. Eskridge, Jr., *The Many Faces of Sexual Consent*, 37 WM. & MARY L. REV. 47 (1995); Heidi M. Hurd, *Was the Frog Prince Sexually Molested?: A Review of Peter Westen's* The Logic of Consent, 103 MICH. L. REV. 1329 (2005) (book review); Stephen J. Schulhofer, *Consent: What It Means and Why It's Time to Require It*, 47 U. PAC. L. REV. 665 (2016).

[266] *Ex Parte* Jones, No. 12-17-00346-CR, 2018 WL 2228888, at *7 (Tex. Ct. App. May 16, 2018).

[267] *See* H.B. 98, 86th Leg. Sess. (Tex. 2019).

[268] State v. VanBuren, 214 A.3d 791, 812 (Vt. 2019) (citing VT. STAT. ANN. tit. 13, § 2606(b)(1), (2) (West 2015)).

[269] *See* Dripps, *supra* note 82, at 975; Tuerkheimer, *supra* note 82, at 4; Falk, *supra* note 82, at 286.

[270] *See* Commonwealth v. Mlinarich, 498 A.2d 395, 402 (Pa. Super. Ct. 1985); Dripps, *supra* note 82, at 975.

C. *Taking the Fight to the Algorithm*

For different reasons, the initial act of wrongful distribution is often of little practical import to the fight against nonconsensual pornography. At times, the sexual content is illegally obtained and distributed by unknown parties who enjoy online anonymity. Even when wrongdoers are identifiable, they can be judgment proof: devoid of any assets that could even begin to compensate the victims for the harms they suffer. Although criminal law purports to overcome this challenge, criminal deterrence often does very little to prevent crimes, instead being an instrument for communicating society's condemnation of the act ex post facto.[271]

Once their images have been made available online, victims are mainly focused on taking them down and preventing their further distribution to the best of their abilities. At times, these efforts involve taking on unsympathetic or outright malicious websites that thrive on victims' plights. However, often, victims suffer the most harm from images distributed on mainstream platforms. Most people, it can only be hoped, do not frequent websites dedicated to nonconsensual pornography in search of their acquaintances; however, the involvement of social media platforms can connect the images with an identifiable and familiar person, exposing victims before their entire social world, workplace contacts, and family members.[272]

Luckily, sustained advocacy efforts have, over time, moved platforms to acknowledge their pivotal role in the fight against nonconsensual pornography. Since 2015, major social platforms, including Reddit, Facebook, Twitter, Google, and Snapchat, have banned nonconsensual pornography and implemented complaint and takedown procedures to assist victims.[273] In addition to responding to complaints regarding specific images, platforms have used hashing technology and matching algorithms to detect and remove any copies circulating within their networks.[274]

Still, once an image is uploaded to a network, it is practically impossible to control its spread outside it.[275] As a consequence, the only technical measure capable of responding to the enduring harms of unwanted distribution is proactive detection before images are

[271] *See* Danielle Keats Citron, *Law's Expressive Value in Combating Cyber Gender Harassment*, 108 MICH. L. REV. 373 (2009).

[272] *See* Solon, *supra* note 36.

[273] *See id.*; Franks, *supra* note 240, at 1270–71.

[274] *See supra* notes 180–182 and accompanying text.

[275] Although platforms at times prevent the downloading of materials, there are easily available technical means of circumventing such obstacles.

distributed.[276] As Mary Anne Franks, one of the leaders of the fight against nonconsensual pornography, reports, companies were initially dismissive of this approach.[277] Nevertheless, in 2017 Facebook launched a pilot program meant to make its takedown efforts proactive by inviting users who worry that their images would be distributed to preemptively send them to Facebook to be hashed, preventing their upload to the network. However, Facebook's problematic track record with respecting user privacy, as well as the very limited applicability of this approach, resulted in much ridicule and very little effect.[278]

Then, in March 2019, Facebook announced that it had put in place algorithmic measures that use machine learning to independently detect nonconsensual sexual content *upon upload*. As Davis described in her announcement, content flagged by the algorithm is reviewed by a "specially-trained member of [Facebook's] Community Operations team"; if the material is found to violate Facebook's prohibition on nonconsensual distribution, it is removed, and in most cases the uploader's account is disabled, subject to an appeal process.[279]

Although Facebook has shared few details about the operation of this system, Facebook employees told news outlets that the filtering algorithm is trained to develop a model of nonconsensual pornography involving "many signals" that presumably indicate "whether an intimate or nude image or video is shared without someone's consent."[280] To supply the learning algorithm with the sufficiently large amount of labeled training data it requires, Facebook turned to the only readily available source of such information: past decisions made in response to takedown requests.[281]

As suggested in Section II.C, Davis is correct to describe the turn to preemptive filtering as the "next frontier" of content moderation.[282] The importance of the shift is not just that it is a more effective form of

[276] *See* Franks, *supra* note 240, at 1273; Eric Goldman, *The Sex Tape Problem . . . and a Possible Legislative Solution?*, TECH. & MKTG. L. BLOG (July 11, 2008), https://blog.ericgoldman.org/archives/2008/07/the_sex_tape_pr.htm [https://perma.cc/UAT4-G6PA].

[277] Franks, *supra* note 240, at 1272–74.

[278] *See NCII Pilot*, FACEBOOK, https://www.facebook.com/safety/notwithoutmyconsent/pilot [https://perma.cc/4ARP-48SK]; Travis M. Andrews, *Facebook Says It Needs Your Explicit Photos to Combat Revenge Porn*, WASH. POST (Nov. 8, 2017), https://www.washingtonpost.com/news/morning-mix/wp/2017/11/08/facebook-says-its-needs-your-explicit-photos-to-combat-revenge-porn [https://perma.cc/3J6L-YZXQ].

[279] Davis, *supra* note 2.

[280] Solon, *supra* note 36; Melanie Ehrenkranz, *Facebook Says It Will Use AI to Police Revenge Porn, but It Won't Fully Explain How (Updated)*, GIZMODO (Mar. 15, 2019, 1:20 PM), https://gizmodo.com/facebook-needs-to-better-explain-how-its-going-to-use-a-1833323427 [https://perma.cc/YJ5Q-RV85].

[281] *See* Solon, *supra* note 36.

[282] *Id.*

content moderation but also that the measure of relevance used by such algorithms changes from factual similarity to the information's legal classification as prohibited distribution. As described above, reliance on such algorithms can have two troubling results: it can, as described in Part I, affect the meaning of nonconsent, and it can, as the filter bubble theory suggests, cement nonconsent as the line that separates acceptable from violative distribution.

The first of these concerns continues the line of traditional criticisms of algorithmic decision-making.[283] As the system's designers disclose, the learning algorithm models nonconsent partly by tracking language and other signals that can identify the uploader's malicious intent.[284] As such external features can be more easily ascertainable than the victim's mental disposition, there is a risk that the algorithm will overemphasize the malicious intent component in constructing its model of nonconsent. As described in Section I.D, this construction can in turn shape how content moderators come to view the meaning of nonconsensuality and, as these pivotal decisions shape the legal discourse that responds to them, the very meaning of this offense, regardless of what meaning legislators intended for it to have.[285]

And, as the filter bubble theory suggests, using a filtering algorithm that uses nonconsent as its measure of relevance can also limit the normative discourse surrounding violative distribution to nonconsensual cases. The algorithm can have this control over the shape of the normative discourse both by omitting from the discussion "irrelevant" information and by legitimizing consensual harms by deeming them irrelevant to the discussion.

The contribution any single case of unwanted distribution of sexual content has to the normative discourse is thus determined by the algorithm's response to the perpetrator's attempt to upload it. If the algorithm finds the upload to be nonconsensual, human content moderators are notified, potentially obligating them to inform the authorities and the victims.[286] Hence, as social media platforms become the focal points for the fight against unwanted distribution, algorithms directly control which cases give rise to legal consequences and get to shape the meaning of violative distribution. As the algorithm measures each case's relevance according to its nonconsensuality, the resulting

[283] *See supra* Section I.D.

[284] Solon, *supra* note 36.

[285] *See* Coglianese & Lehr, *supra* note 94, at 1218.

[286] Notably, the CDA does not give platforms immunity from prosecution for federal crimes that can give rise to reporting duties, such as terrorism and child abuse. *See* 47 U.S.C. § 230(e)(1); *cf.* Tsesis, *supra* note 3.

filter bubble will tether the subsequent normative discourse to this single legal category.

Conversely, if the algorithm determines that the upload is consensual, the sexual content would go on to be distributed, subject to other restrictions.[287] If those harmed by the distribution seek to act against it, they will face an uphill battle to change the minds of content moderators accustomed to addressing only nonconsensual violations and unskilled in dealing with other harms,[288] and to affect a desensitized public opinion accustomed to viewing consensual distribution as indistinguishable from nonvolatile pornography. Ironically, by invisibly hiding consensual harms from content moderators, the effects of algorithmic filtering can be that such harms are hidden in plain sight for the rest of society, as it increasingly relies on platforms' judgment to mark the boundaries between the acceptable and unacceptable in the virtual domain.[289]

D. *The Relevance of Nonconsent*

There may be good reasons to limit the immediate legal meaning of violative distribution to nonconsensual distribution. Still, as discussed below, legal decision-making and legal discourse cannot be thus limited, lest they grow indifferent to the existence of consensual harms and ignore the very need for normative discussion.

1. The Harms of Unwanted Distribution

Much has been written on the terrible and enduring harms wrought by the nonconsensual distribution of one's sexual images.[290] As a consequence of violative distribution, many victims have suffered

[287] It should be noted that Facebook currently screens for images that contain nudity in general. Other platforms, however, do not ban explicit images but are nonetheless endeavoring to prevent nonconsensual distribution. *See Non-consensual Nudity Policy*, TWITTER (Nov. 2019), https://help.twitter.com/en/rules-and-policies/intimate-media [https://perma.cc/PAB9-DSKW]; *Nudity and Sexual Content Policies*, YOUTUBE, https://support.google.com/youtube/answer/2802002?hl=en-GB [https://perma.cc/JPH7-GF8P]; *Reddit Content Policy*, REDDIT, https://www.redditinc.com/policies/content-policy [https://perma.cc/3ZUX-ERPW].

[288] *See* Shannon Vallor, *Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character*, 28 PHIL. & TECH. 107 (2015). I thank Christina Spiegel for pointing out the de-skilling effect of such filtering.

[289] As recalled, this legitimizing effect stood at the backdrop of *Stratton Oakmont, Inc. v. Prodigy Services Co.* that had led to the enactment of the CDA. Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

[290] *See* Franks, *supra* note 240, at 1273.

direct emotional trauma from knowing that they have lost control over their sexual image, suffered from loss of employment opportunities and borne other economic costs as potential employers and commercial associates negatively reacted to the availability of the images, suffered harm to their ability to trust others and to form intimate relationships, experienced fear of public appearances as they worry whether people they encounter have viewed their images, and many other related emotional, social, and relational harms.[291] These harms have been known to produce severe psychological trauma, depression, and eating disorders, and have even driven some victims to take their own lives.[292] In addition to these reputational harms,[293] distributing a person's sexual images against their will is a grave violation of their sexual autonomy, squashing their ability to control the scope of their sexual exposure to others and their ability to choose with whom to interact sexually.[294]

All these harms, however, do not necessarily result from the *nonconsensuality* of the distribution, or at least not solely so. Think, for instance, of the *Hustler* cases: despite the courts' holdings, much of the harm done to the victims was caused not by their being misrepresented as consenting to the publication but by the publication itself. While it is likely that the courts stressed the publications' nonconsensuality to avoid intruding on the magazine's constitutionally protected right to publish consensual pornography, this focus certainly creates a distorted image of the harms at stake. The emotional, occupational, and reputational harms suffered by the victims of nonconsensual distribution can likewise affect those who did not want the distribution of their images but formally consented to it due to material need or emotional duress or for the host of other reasons that can lead people to participate in consensual but unwanted sexual interactions.[295]

[291] *See* People v. Austin, 155 N.E.3d 439, 469 (Ill. 2019).

[292] *See* Citron & Franks, *supra* note 243, at 350–51.

[293] These harms are reputational in the sense that they result from strangers and acquaintances having access to the victim's sexual images and the victim's knowledge of this. *See* Goldman, *supra* note 276. As Eric Goldman comments, the reputational harms that result from nonconsensual distribution could abate if people stopped seeing the distribution of one's sexual images as something having a negative reputational effect. Eric Goldman, *What Should We Do About Revenge Porn Sites like Texxxan?*, FORBES (Jan. 28, 2013, 1:13 PM), https://www.forbes.com/sites/ericgoldman/2013/01/28/what-should-we-do-about-revenge-porn-sites-like-texxxan/?sh=75e69b7eff8d [https://perma.cc/22DW-9J7M].

[294] *See* Rosenberg & Dancig-Rosenberg, *supra* note 241.

[295] As Robin West puts it:

> Heterosexual women and girls, married or not, consent to a good bit of unwanted sex with men that they patently don't desire, from hook-ups to dates to boyfriends to co-habitators, to avoid a hassle or a foul mood the endurance of which wouldn't be worth

The same holds true for the damage done by unwanted distribution to the individual's sexual autonomy. Sexual violations are not limited to unwanted physical contact, just as sexuality itself is not confined to sexual intercourse.[296] Nor does sexual autonomy begin and end with legal consent. Real-world violations of sexual autonomy range from the archetypical but statistically negligible stranger rape to much more common acquaintance and intimate partner sexual assaults and various forms of nonphysical duress.[297] Virtual violations likewise exist on a spectrum that ranges from stranger hacking to intimate betrayals of trust and acquiesced-to but unwanted distribution. On both spectrums, all violations can be detrimental to one's sexual autonomy and well-being, even though not all should or can be legally prohibited.

Like physical sexuality, the distribution of one's sexual images can have a positive and liberating effect on one's sexual well-being and autonomy, but not when one merely or formally consents to the distribution without wanting it. As India Thusi notes, the rapid increase in nonprofessional pornography promises to remove exploitation from the pornography industry by giving creators complete control over the production and distribution of their sexual images.[298] However, as Mary Anne Franks realistically puts it, "The problem is that a good thing can't exist for more than two seconds before someone comes along and makes it a horrible thing."[299] In some cases, persons consent to the distribution without fully realizing the effect of having their sexual

> the effort, to ensure their own or their children's financial security, to lessen the risk of future physical attacks, to garner their peers' approval, to win the approval of a high status man or boy, to earn a paycheck or a promotion or an undeserved A on a college paper, to feed a drug habit, to survive, or to smooth troubled domestic waters. Women and girls do so from motives of self-aggrandizement, from an instinct for survival, out of concern for their children, from simple altruism, friendship or love, or because they have been taught to do so. But whatever the reason, some women and girls have a good bit of sex a good bit of the time that they patently do not desire.

Robin West, *Sex, Law, and Consent*, *in* THE ETHICS OF CONSENT: THEORY AND PRACTICE 221, 236 (Franklin G. Miller & Alan Wertheimer eds., 2010).

[296] *See supra* Section III.B; *see also* Nicola Lacey, *Unspeakable Subjects, Impossible Rights: Sexuality, Integrity and Criminal Law*, 11 CANADIAN J.L. & JURIS. 47, 65–66 (1998). The move from the physical to the symbolic reflects a broader view of sexuality that is detached from its biological origins. The origins of this image, William Eskridge, Jr. suggests, and the notion that the "verbal and physical drama" that surrounds the physical act "is what makes sex 'sexy,'" can be traced back to the inclusion of gay sexuality in public discourse. Eskridge, *supra* note 265, at 63.

[297] *See* Maggen, *supra* note 24, at 611.

[298] Thusi, *supra* note 236.

[299] Rebecca Hiscott, *Why Amateur Porn Will Never Be Safe*, MASHABLE (Jan. 12, 2014), https://mashable.com/archive/amateur-porn-abuse [https://perma.cc/D766-W7M8] (quoting Mary Anne Franks).

images distributed online.[300] Others erroneously assume that distribution can be limited in its scope.[301] Such people realize only in hindsight that the internet is without boundaries and it never forgets, and that once images are distributed, they have lost all control over the spread of their sexual images, which are made available to friends, colleagues, family members, and potential employers.[302]

Often in such instances, persons are pressured into accepting the creation and distribution of their images. At times this is a result of emotional duress within an abusive intimate relationship.[303] Much more often, nonprofessional performers can succumb to growing material and psychological duress and consent to the creation and distribution of increasingly explicit materials.[304] In all these cases, once the materials are made available online, there is no turning back, and there is little that can be done to prevent the potential effects of the distribution on their lives, livelihood, and sexual well-being.[305]

All this is not to say that all unwanted distribution of sexual imagery should be prohibited, but it does suggest that, as is the case with consensual but unwanted sexual contact, we cannot blind ourselves or become callous to the existence of such harms.[306] Nevertheless, this is precisely what could happen as a result of a legalistic filter bubble that subjects only formally nonconsensual distribution to human scrutiny.

[300] *See, e.g.*, Lane v. MRA Holdings, LLC, 242 F. Supp. 2d 1205 (M.D. Fla. 2002).

[301] *See* Harriet Grant, *Group of US Women Sue "Amateur" Porn Producer over "Coercion and Lies*,*"* GUARDIAN (Sept. 20, 2019, 8:22 AM), https://www.theguardian.com/world/2019/sep/20/us-women-sue-porn-producer-over-alleged-coercion-and-lies [https://perma.cc/64B8-H8PF]; Evelyn Nieves, *A Festival with Nudity Sues a Sex Web Site*, N.Y. TIMES. (July 5, 2002), https://www.nytimes.com/2002/07/05/us/a-festival-with-nudity-sues-a-sex-web-site.html [https://perma.cc/2BKW-44M7].

[302] *See* Grant, *supra* note 301. Even professional models who willingly consent to the creation of sexual materials for commercial purposes can be horrified to learn how little control they subsequently have over their sexual images. *See* Emily Ratajkowski, *Buying Myself Back: When Does a Model Own Her Own Image?*, N.Y. MAG.: THE CUT (Sept. 15, 2020), https://www.thecut.com/article/emily-ratajkowski-owning-my-image-essay.html#_ga=2.29517823.1729852367.1624048196-1477425549.1624048196 (last visited Apr. 23, 2022).

[303] *See* Hiscott, *supra* note 299.

[304] For a powerful depiction of this dynamic, see HOT GIRLS WANTED (Two to Tangle Productions 2015).

[305] *See* Farley, Franzblau & Kennedy, *supra* note 237, at 1079.

[306] *See* Robin L. West, *Legitimating the Illegitimate: A Comment on Beyond Rape*, 93 COLUM. L. REV. 1442 (1993); *see also* Donald A. Dripps, *More on Distinguishing Sex, Sexual Expropriation, and Sexual Assault: A Reply to Professor West*, 93 COLUM. L. REV. 1460 (1993).

2. The Debate on Sexual Autonomy

As algorithmic filtering singles out nonconsensual violations and deems all other violations irrelevant, it reduces the number of abusive but formally consensual cases that reach human decision-makers, and at the same time legitimizes them by giving them an implicit seal of approval so that they become indistinguishable from otherwise protected and even celebrated pornography.[307] Both effects can hinder the development of normative debate on the place of formal consent in the assessment of a distribution's social acceptability.

Such stifled debate would stand in opposition to the vibrant discourse surrounding the meaning of consent in real-world sexual violations, as manifested in the #MeToo movement.[308] This debate often involves two opposing views on the meaning of sexual autonomy, which stress either the transactional nature of personal autonomy or sexuality's idiosyncrasy.[309] To the former, dominant school of thought, sexual autonomy is essentially synonymous with formal consent, in its general legal meaning, and should be protected against forms of coercion generally held to vitiate legal consent.[310] According to this view, in protecting sexual autonomy, law's task is to secure it the same protections that facilitate the free flow of commodities in the free market.[311]

[307] *Cf.* Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

[308] *See, e.g.*, Tristin K. Green, *Was Sexual Harassment Law a Mistake? The Stories We Tell*, 128 YALE L.J.F. 152, 154 (2018); Melissa Murray, *Consequential Sex: #MeToo,* Masterpiece Cakeshop, *and Private Sexual Regulation*, 113 NW. U. L. REV. 825, 833 (2019); Joan C. Williams et al., *What's Reasonable Now? Sexual Harassment Law After the Norm Cascade*, 2019 MICH. ST. L. REV. 139, 152.

[309] As Stephen Schulhofer puts it, "[T]he major disagreement on this issue is between those who want the list to be very short—limited to things that are almost as coercive as physical violence—and on the other side, those who want that list to include many or all the other circumstances that limit a completely free choice." Stephen J. Schulhofer, *Reforming the Law of Rape*, 35 LAW & INEQ. 335, 345 (2017); *see also* Deborah Tuerkheimer, *Slutwalking in the Shadow of the Law*, 98 MINN. L. REV. 1453, 1490 (2014); Robin West, *Law's Emotions*, 19 RICH. J.L. & PUB. INT. 339, 349 (2016).

[310] *See, e.g.*, Dripps, *supra* note 265, at 1791; Dsouza, *supra* note 265, at 520–21.

[311] As Donald Dripps suggests, "[S]exual cooperation is a service much like any other, which individuals have a right to offer for compensation, or not, as they choose. Consequently, sexual autonomy means freedom from illegitimate pressures to provide this particular service." Dripps, *supra* note 265, at 1786 (footnotes omitted); *see also* David P. Bryden, *Redefining Rape*, 3 BUFF. CRIM. L. REV. 317, 445 (2000); RICHARD A. POSNER, SEX AND REASON 384–95 (1992); Dripps, *supra* note 265, at 1791–92; Schulhofer, *supra* note 309, at 346–47. Margaret Radin describes this view as "universal commodification." *See* Margaret Jane Radin, *Market-Inalienability*, 100 HARV. L. REV. 1849, 1859–70 (1987); *see also* Laina Y. Bay-Cheng & Rebecca K. Eliseo-Arras, *The*

Opposing this view are those who believe that formal consent and physical coercion cover only part of what constitutes a wrongful violation of sexual autonomy.[312] The emphasis on formal consent, this view suggests, is appropriate for a transactional setting in which people's interests stand opposite to each other but is utterly foreign to the sexual domain, which is grounded in mutuality.[313] Martha Chamallas, one of the first to clearly introduce mutuality in opposition to the consent paradigm, describes it as an egalitarian stance meant "to afford women the power to form and maintain noncoercive sexual relationships, both within and outside of marriage."[314] Robin West, one of the leading voices in the egalitarian camp, describes the harm caused by exploitative sexual relations as a self-alienating condition of "sexual dysphoria," in which the victim's sexuality is denied the special place it commonly enjoys in our emotional lives.[315] The harms of such alienation are not only emotional or psychological but also political, reducing the victim's "instincts and desire for social, sexual, and commercial connection with others, to a series of permissions borne of precious little but shrunken visions, sour grapes, and material necessity."[316] As West describes it, using material, emotional, and other forms of compulsion to pressure another individual into an unwanted sexual situation pits their rational self against their hedonic self, leading to the erosion of the victim's distinct sexual personhood.[317]

The egalitarian approach does not necessarily suggest that the harms caused by unwanted sexual relations ought to be the subject of criminal prohibition.[318] It does, however, demand that we acknowledge their place within the normative discussion on sexual wrongdoing.[319] Exclusively focusing on the legalistic category of consent, this view argues, can suppress this inclusive discussion.[320] Indeed, as the #MeToo

Making of Unwanted Sex: Gendered and Neoliberal Norms in College Women's Unwanted Sexual Experiences, 45 J. SEX RSCH. 386, 395 (2008).

[312] *See, e.g.*, JOHN GARDNER, *The Wrongness of Rape*, *in* OFFENCES AND DEFENCES 1, 17 (2007).

[313] *See, e.g.*, Catharine A. MacKinnon, *Rape Redefined*, 10 HARV. L. & POL'Y REV. 431, 441 (2016); Tuerkheimer, *supra* note 309, at 1476.

[314] Martha Chamallas, *Consent, Equality, and the Legal Control of Sexual Conduct*, 61 S. CAL. L. REV. 777, 783 (1988).

[315] Robin West, *Consensual Sexual Dysphoria: A Challenge for Campus Life*, 66 J. LEGAL EDUC. 804 (2017).

[316] West, *supra* note 309, at 350; West, *supra* note 315, at 808.

[317] West, *supra* note 315, at 811; *see also* JANET HALLEY, SPLIT DECISIONS: HOW AND WHY TO TAKE A BREAK FROM FEMINISM 63 (2006).

[318] *See, e.g.*, Tuerkheimer, *supra* note 309, at 1490; West, *supra* note 309, at 349; West, *supra* note 306.

[319] Maggen, *supra* note 24, at 610–15; *see also* West, *supra* note 315, at 806.

[320] *See, e.g.*, Tuerkheimer, *supra* note 309, at 1490; West, *supra* note 309, at 349; West, *supra* note 306.

movement demonstrated, the social attitude toward a violation of sexual autonomy can be no less important than its legal meaning, and strict adherence to legal categories can be used to avoid a social discussion.[321] As Kat Stoeffel puts it, "[I]t seems like every time someone explains that women and men do not always meet for sex on equal footing, the conversation collapses into a black-and-white debate of Was It Rape."[322] Only by going beyond this strictly legal distinction can we have a vigorous discussion on its appropriateness.

The introduction of legalistic filter bubbles can, however, turn nonconsent from merely a conversation stopper to an axiomatic precondition, preventing a conversation on the place of consent from ever taking place. Even though it is unlikely that the result of filtering will be the complete obfuscation of consensual harms, it is likely that it will significantly reduce the occasions in which such cases will reach decision-makers and broader public attention. When they do, the conditioning caused by the legalistic filtering is likely to move content moderators and society to view them as unfortunate but normatively irrelevant to the free exercise of sexual autonomy—trivializing them just as many forms of sexual abuse were in the past.[323] The few exceptional cases that manage to survive this gauntlet will scarcely amount to the critical mass that often provokes significant normative discussion.[324]

[321] *See, e.g.*, Sarah K. Burgess, *Between the Desire for Law and the Law of Desire: #MeToo and the Cost of Telling the Truth Today*, 51 PHIL. & RHETORIC 342, 346 (2018) ("For advocates of the movement, #MeToo operates in extralegal spaces to define and negotiate what the laws of desire should be."); Fiona Chen, *Why the Aziz Ansari Story and Discussions of Grey Areas Are Central to the #MeToo Movement*, TECH (Jan. 25, 2018), https://thetech.com/2018/01/25/me-too-aziz-ansari [https://perma.cc/JM5Y-Q5BG] ("For now, our focus should be on socially transforming the way we understand sexual violence."). *But see* Murray, *supra* note 308, at 873 ("[T]he #MeToo movement's actions are not simply about usurping the state's regulatory role and imposing consequences on those who have failed to comply with the movement's understanding of appropriate sexual conduct. Instead, the larger goal is to persuade the state to adopt this vision of appropriate sex and sexuality and use it to undergird more progressive and egalitarian laws and policies.").

[322] Kat Stoeffel, *It Doesn't Have to Be Rape to Suck*, N.Y. MAG.: THE CUT (Oct. 6, 2014), https://www.thecut.com/2014/10/doesnt-have-to-be-rape-to-suck.html (last visited Apr. 23, 2022) (discussing the precursors to #MeToo in relation to the inadequacy of rape law); *see also* Kimberly Kessler Ferzan, *Consent and Coercion*, 50 ARIZ. ST. L.J. 951, 1007 (2018) ("We ought to be able to have conversations about how people treat each other, and the terms of sexual negotiations, without a conclusion that the crime is rape.").

[323] To borrow Anupam Chander's term, this would essentially make them legally unprotected "youthful indiscretions." Anupam Chander, *Youthful Indiscretion in an Internet Age*, *in* THE OFFENSIVE INTERNET: SPEECH, PRIVACY, AND REPUTATION 124, 124–25 (Saul Levmore & Martha C. Nussbaum eds., 2010).

[324] *See* Eileen Oak, *A Minority Report for Social Work? The Predictive Risk Model (PRM) and the Tuituia Assessment Framework in Addressing the Needs of New Zealand's Vulnerable Children*, 46 BRIT. J. SOC. WORK 1208, 1215 (2016).

Conclusion

Legalistic filtering, I have argued, is a novel phenomenon capable of having far-reaching consequences. It differs from past uses of algorithms to assist legal decision-making in that it uses legal analysis to determine what information is brought before human decision-makers—what cases they get to decide. The rise of legalizing filtering, I suggested, is prone to creating legalistic filter bubbles, in which normative debate is limited in scope to the constraints set by the legal classification used in the filtering process.

In the case of the fight against violative distribution of sexual materials, the emergence of a filter bubble can effectively limit the legal and social meaning to prohibited distribution to nonconsensual distribution, concealing the existence of consensual harms and essentializing consent as the sole legal measure of sexual autonomy. This effect, I suggested, would stymie the development of a vibrant discourse on the meaning of virtual sexual autonomy akin to the discussion currently surrounding physical sexuality.

Although there may be good, even overwhelming, reasons to accept nonconsent as the sole legal measure of wrongdoing, legal discourse that is without vibrant debate can only grow moribund. Robert Cover famously described this dynamic as law's "jurispath[y]"; law requires ongoing revitalization through encounters with external, jurisgenerative narratives to remain vital.[325] Law, as it appears in the paradigms dominating positive law, is but a part of the greater normative universe law inhabits; although it is self-sufficient for legal analysis, the practice of legal adjudication cannot be normatively complete without the inclusion of extralegal narratives.[326] By restricting legal discourse to that which is legally relevant, legalistic filter bubbles foster this already deeply rooted jurispathic tendency to disregard narratives that oppose prevailing legal paradigms.

[325] *See generally* Cover, *supra* note 20.

[326] *See id.* at 4–6.