#MeTooBots and the AI Workplace

by

Leora Eisenstadt
Fox School of Business – Temple University

Abstract

Responding to the #MeToo Movement, companies across the United States and Europe are beginning to offer products that use AI to detect discrimination and harassment in digital communications. These companies promise to outsource a large component of the EEO compliance function to technology, preventing the financial costs of toxic behavior by using AI to monitor communications and report anything deemed inappropriate to employer representatives for investigation. Highlighting the problem of underreporting of sexual harassment and positing that many victims do not come forward out of a fear of retaliation, these "#MeTooBots" propose to remove the human element from reporting and rely on AI to detect and report unacceptable conduct before it contaminates the workplace.

This new technology raises numerous legal and ethical questions relating to both the effectiveness of the technology and the ways in which it alters the paradigm on which anti-discrimination and anti-harassment doctrine is based. First, the notion that AI is capable of identifying and parsing the nuances of human interactions is problematic as are the implications for underrepresented groups if their linguistic styles are not part of the AI's training. More complicated, however, are the questions that arise from the technology's attempt to eliminate the human reporter: (1) How does the use of AI to detect harassment impact employer liability and available defenses since the doctrine has long been based on worker reports? (2) How does this technology impact alleged victims' vulnerability to retaliation when incidents may be detected without a victim's report? (3) What is the impact on the power of victim voice and autonomy in this system? and (4) What are the overall consequences for organizational culture when this type of technology is employed?

This Article examines the use of AI in EEO compliance and considers whether the elimination of human reporting requires a reconsideration of the U.S.'s approach to discrimination and harassment. Appearing on the heels of revelations about the use of non-disclosure agreements and arbitration clauses to silence victims of sexual harassment, this Article posits that the use of AI to detect and report improper communications, an innovation that purports to help eradicate workplace harassment, may, in reality, be problematic for employers and employees alike, including functioning as a new form of victim abuse. Lastly, the Article considers the difficult work of creating open, healthy workplace cultures that encourage reporting, and the impact of outsourcing this work to Artificial Intelligence. Rather than rejecting what may be an inevitable move towards incorporating artificial intelligence solutions in the workplace, this Article suggests more productive uses of AI at work and adjustments to employment discrimination doctrine to be better prepared for an AI-dependent world.

Introduction

In January 2020, *The Guardian* published a short article describing the development of Artificial Intelligence (AI) solutions to the problem of sexual harassment in the workplace.[1] Referred to as "#MeTooBots," the AI-infused technology can reportedly "monitor and flag communications between colleagues and [is] being introduced by companies around the world."[2] The companies offering this technology explain its development as a response to the staggering number of women who experience sexual harassment in the workplace, the cost to employees, employers, and workplace culture overall, and most importantly, the fact that sexual harassment is massively underreported largely because victims fear retaliation if they report.[3] As one company explains:

> By leveraging AI-infused technology, . . . organizations can identify, investigate, and handle offensive communications in the early stage -- without requiring the victim to report the incident to a superior. With artificial intelligence in the workplace, the 75% of sexual harassment cases that typically go unreported, can be automatically identified. Armed with this technology, organizations can protect employees, the company, and the culture from malicious employees who would otherwise be toxic to the workforce.[4]

This technology is under development by multiple companies but the idea is generally the same—"the bot uses an algorithm trained to identify potential bullying, including sexual harassment, in company documents, emails and chat. . . . with anything the AI reads as being potentially problematic then sent to a lawyer or HR manager to investigate."[5] AI scientists have voiced some skepticism, noting that the idea was promising but perhaps limited in its current capabilities.[6] These critiques center on the nuanced nature of human communication and the notion that AI is not yet capable of understanding the subtleties and complexities inherent in harassing language.[7] Other critics focus on the privacy implications and the "Big Brother" quality, arguing that it would be viewed by workers as "another way [for employers] to control their employees."[8] One commentator called this approach "an Orwellian misuse of AI," expressing particular concern for employees accused of harassment by the bot. "Any risqué joke, comment on appearance, proposal to go out for drinks, or even the stray mention of a body part will probably be meticulously logged to be used against you at a future date."[9]

The urge to simply write off this new technology as ineffective or immature is a mistake as is the narrow critique that its essential ethical and legal problems lie in privacy concerns or employer control. In fact, #MeTooBots emerge from a workplace that is speeding towards incorporation of AI in numerous forms and functions.[10] The fact that the technology may not be quite capable of

---

[1] Isabel Woodford, *Rise of #MeTooBots: scientists develop AI to detect harassment in emails*, THE GUARDIAN, Jan. 3, 2020, https://www.theguardian.com/technology/2020/jan/03/metoobots-scientists-develop-ai-detect-harassment.
[2] Woodford, *supra* note 1.
[3] *See, e.g.,* AWARE, HTTP://WWW.AWAREHQ.COM/BLOG/IDENTIFYING-AND-REDUCING-WORKPLACE-SEXUAL-HARASSMENT-WITH-AI (last visited May 18, 2021).
[4] *Id.*
[5] Woodford, *supra* note 1.
[6] *Id.*
[7] *Id.*
[8] *Id.*
[9] Norman Lewis, *#MeTooBots that will scan your personal emails for 'harassment' are an Orwellian misuse of AI*, RT (Jan. 3, 2020), https://www.rt.com/op-ed/477393-metoobots-ai-orwellian-harassment/.
[10] *See infra* Part I.A.

the nuanced task it is being given will likely be seen by some as a temporary problem and by others as no problem at all. And the exclusive focus on privacy and the "big brother" aspects of this technology misses key ethical and legal problems in its adoption. As this Article describes, the embrace of #MeTooBots creates unforeseen problems for the wrongly accused, the employer, and the victim. Basic problems with the technology may end up exposing women and people of color to unwarranted investigations, creating a more hostile workplace for already marginalized employees. At the same time, use of AI as proposed by the companies offering these tools may also open employers to significantly increased legal liability. Lastly, assigning this function to AI has the potential to expose victims of digital harassment to retaliation without legal protection.[11] Ironically, use of AI in this space is likely to have the opposite of its intended effect. Rather than serving the purpose of creating healthier workplace cultures by identifying unreported sexual harassment, AI sexual harassment monitors would likely generate greater distrust among employees and managers alike, subject victims of harassment to lawful retaliation, and perhaps most profoundly, impact the functioning of employment law doctrine in this area.

The use of AI in place of victim reporting is a monumental shift in the expectations on which discrimination and harassment doctrine is based. It threatens to weaken victim voice and agency within the system. Even more alarming, it has the potential to alter the basic premise on which Title VII and its doctrine rests—that the law lays out basic prohibitions but relies on victims or other human reporters to bring violations to the attention of employers, regulators, and courts. In fact, the determination of employer liability, the prohibition on retaliation in anti-discrimination laws, and the role and function of the Equal Employment Opportunity Commission (EEOC) are all based on the expectation that human reporters are integral to the system.[12] There are no discrimination police—the system is based on *people* coming forward to report instances of unlawful discrimination and the protection of those reporters. Were the technology and approach of #MeTooBots to be adopted broadly, it would necessitate a rethinking of employment discrimination doctrine, the role of victims and their agency in a process that advances without their control, and how retaliation protection should function in such a system.

This Article considers the use of AI in detecting and reporting sexual harassment in the context of an overwhelming movement to incorporate technology and AI in the workplace—in hiring, worker assessment and tracking, and worker and workplace optimization.[13] #MeTooBots are not a one-off creation but rather one in a series of technological developments that are quickly being normalized. Although generally referred to by the companies that produce them as "AI" or AI-based solutions,[14] some scholars have noted that AI is defined differently in legal literature,

[11] *See infra* Parts III. B. 1 & 2.

[12] It is important to note here that many scholars view retaliation as a form of discrimination itself. See Deborah Brake, *Retaliation*, 90 MINN. L. REV. 18, 21 (2005) (contending that "[r]ecognizing retaliation as a form of discrimination, one that is implicitly banned by general proscriptions of discrimination, pushes the boundaries of dominant understandings of discrimination in useful and productive ways."); see also Brief of Employment Law Professors as Amici Curiae in Support of Respondent at 5, Univ. of Tex. Southwestern Med. Ctr. v. Nassar, 133 S. Ct. 2517 (2013) (No. 12-484) (contending "[a] long line of cases confirms that when Congress uses the word "discriminate" that term encompasses retaliation."). *See also* Pauline Kim, *Panel V: Proving Discrimination: Addressing Systemic Discrimination: Public Enforcement and the Role of the EEOC*, 95 B.U.L. REV. 1133, 1137-38 (describing the original and evolving role of the EEOC including investigation and enforcement of Title VII).

[13] *See* Richard A. Bales and Katherine V.W. Stone, *The Invisible Web at Work: Artificial Intelligence and Electronic Surveillance, Under the Labor Laws*, 41 BERKELEY J. EMP. & LAB. L. 1, 9-21 (2020).

[14] *See, e.g.* AWARE, https://www.awarehq.com/blog/identifying-and-reducing-workplace-sexual-harassment-with-ai (last visited May 20, 2021); REVEAL, https://www.revealdata.com/ (last visited May 20, 2021).

technical literature, and in the popular press and industry publications.[15] This Article generally uses the term "AI" because that is how the companies that market so called #MeTooBots refer to them—in fact, regardless of the accuracy of the designation, the notion that AI can replace formerly human tasks appears to be their key selling point.[16] The terms "machine learning" and "cognitive computing" may also be used as they are common labels for the type of technology that encompasses #MeTooBots.

The Article proceeds as follows: Part I considers the environment in which this technology arose—the increasingly digital workplace in which many workers find themselves and the pressures of the #MeToo movement. Part II examines the technology itself, the context for its development, and how it functions. Part III considers the significant issues that arise when such technologies are adopted including the overall effectiveness (or ineffectiveness) of the technology to do what it sets out to do and the resulting impact on underrepresented employees, the impact on employer liability for harassment, the impact on victims of harassment and decreased protection against retaliation, and finally, the impacts on victim voice, organizational culture, and more broadly, employment discrimination law and doctrine. Part IV concludes with some recommendations to two audiences— employers considering adopting this technology and courts applying antidiscrimination law in an AI world—and proposes modifications to both the law and the technology. Rather than rejecting what may be an inevitable move towards incorporating artificial intelligence solutions in the workplace, this Article suggests more productive uses of AI at work and adjustments to employment discrimination doctrine to be better prepared for an AI-dependent world.

I. The Modern Workplace

#MeTooBots and their harnessing of AI to monitor employee communications for harassing or inappropriate words and conduct have developed in the context of two major forces: (1) a world in which, for a large majority of white collar workers, digital communications have wholly replaced in-person interactions with a likelihood that even after the COVID-19 pandemic, many workplaces will never return to an entirely in-person culture[17], and (2) an overwhelming urge in the corporate world to adopt tech solutions to workplace problems.[18] To understand the turn to algorithmic investigation of harassment, it is important to consider the impact of both of these realities.

A. The Remote/Digital Workplace

The global pandemic that hit the world in early 2020 brought innumerable changes to the workplace and the way businesses function. Most prominently, of course, was the transition to remote work for millions of workers as lockdown orders required anyone whose physical presence

[15] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1684 (2020). Ajunwa has chosen to use the terms "algorithms" and "machine learning algorithms" instead of AI, seeing these as more precise technical terms. *Id.*

[16] *See* AWARE, *supra* note 14.

[17] Caroline Castrillon, *This is the Future of Remote Work in 2021*, FORBES (Dec. 27, 2020), https://www.forbes.com/sites/carolinecastrillon/2021/12/27/this-is-the-future-of-remote-work-in-2021/?sh=764635f91e1d (citing a recent CFO survey that "revealed that over two-thirds (74%) plan to permanently shift employees to remote work after the Covid-19 crisis ends.").

[18] *See infra* Part I. A.

at work was not essential to switch to working exclusively from home.[19] A Pew Research Center study of the American workforce conducted in October 2020 found a staggering shift in the way Americans work. Pre-pandemic, approximately 20% of American workers worked from home. Due to COVID-19, 71% of workers now work remotely and 54% want to continue working from home even after the pandemic is over.[20] In addition to this group, approximately one third of American workers want to continue working from home part-time when the pandemic ends, and only 11% would not want to work remotely at all.[21] Perhaps most dramatically, "some 46% of those who rarely or never teleworked before the coronavirus outbreak say they'd want to work from home all or most of the time when the pandemic is over."[22]

Employers are equally on-board with the switch to remote work. "A Gartner, Inc. survey of 317 CFOs and Finance leaders on March 30, 2020 revealed that 74% will move at least 5% of their previously on-site workforce to permanently remote positions post-COVID 19."[23] Some commentators think that these figures are an underestimation. "Companies as disparate as Nationwide Insurance, the 95-year-old financial services giant, and Twitter, a bellwether of the internet economy, have said they will make remote working a much bigger piece of their businesses going forward."[24] This remote work revolution, brought on by the pandemic, is likely to last because it is both a cost-saving measure for employers and seen as a benefit by employees.[25]

This switch to remote work, however, has massive implications for the way workers communicate and for employers' ability to monitor that communication. Before Spring 2020, a large share of work was accomplished via email, messaging, and the like.[26] But the change to remote work has increased that trend dramatically. The notion, oft-seen on Twitter memes, that "this meeting really could have been done by email" was given new life by the lockdowns across the U.S. that necessitated a switch from in-person meetings to a combination of email, messaging, and online conferencing.[27] According to Pew, these "tools have become a vital part of the workday."[28] Eighty-one percent of remote workers "say they use video calling or online conferencing services like Zoom or WebEx to keep in touch with co-workers, with 59% saying

[19] *See* Clive Thompson, *What If Working From Home Goes On … Forever?*, N.Y. TIMES MAGAZINE (June 9, 2020), https://www.nytimes.com/interactive/2020/06/09/magazine/remote-work-covid.html.
[20] Kim Parker, Juliana Menasce Horowitz, and Rachel Minkin, *How the Coronavirus Outbreak Has – and Hasn't – Changed the Way Americans Work,* PEW RESEARCH CENTER (DEC. 9, 2020), https://www.pewresearch.org/social-trends/2020/12/09/how-the-coronavirus-outbreak-has-and-hasnt-changed-the-way-americans-work/.
[21] *Id.*
[22] *Id.*
[23] GARTNER, *Gartner CFO Survey Reveals 74% Intend to Shift Some Employees to Remote Work Permanently* (April 3, 2020), https://www.gartner.com/en/newsroom/press-releases/2020-04-03-gartner-cfo-surey-reveals-74-percent-of-organizations-to-shift-some-employees-to-remote-work-permanently2.
[24] Emily He, *The New Future Of Work In A Post-Pandemic World*, FORBES (June 1, 2020), https://www.forbes.com/sites/emilyhe/2020/06/01/the-new-future-of-work-in-a-post-pandemic-world/?sh=34922ba03382.
[25] Maria Cramer and Mihir Zaveri, *What if You Don't Want to Go Back to the Office?*, N.Y. TIMES (May 5, 2020), https://www.nytimes.com/2020/05/05/business/pandemic-work-from-home-coronavirus.html.
[26] *See* Amy Gallo, *Stop Email Overload*, HARV. BUS. REV. (Feb. 21, 2012) (describing email as both a "threat to productivity" and "an essential work tool.").
[27] *See* Amber Tiffany, *7 Warning Signs Your Meeting Should Be an Email*, GOTOMEETING BLOG (Nov. 2, 2017), https://blog.gotomeeting.com/7-warning-signs-your-meeting-should-be-an-email/
[28] Parker, et. al., *supra* note 20.

they often use these types of services."[29] In addition, "some 57% say they use instant messaging platforms such as Slack or Google Chat at least sometimes (43% use these often)."[30]

With this shift, formal meetings, casual conversations, and even water-cooler chats have morphed into digital communications that leave a record. The most commonly used online conferencing tools like Zoom, Google Meet, Skype, and Microsoft Teams all allow users to record meetings and store them in the cloud or on individual computers.[31] Others provide an automatic transcript of meetings.[32] Similarly, email and internal company messaging systems all leave a digital record of communications.[33] This digital record is searchable, particularly when AI systems that offer the ability to scan through millions of bits of data are unleashed on them.

Of course, the move to increased digital communication still creates opportunities for inappropriate behavior and unlawful harassment. As Edgar Ndjatou, executive director of Workplace Fairness, notes, "you don't have to be in the same place for sexual harassment to happen."[34] In fact, he points out, digital harassment can take any of the following forms:

> Inappropriate comments, jokes, pictures or even GIFs sent via email, chat massages or texts; sexual or discriminatory comments made during video meetings; commenting on a co-worker's appearance during video meetings; emails or text messages requesting sexual favors; stalking on social media; [and] unsolicited and/or inappropriate communications through company messaging apps.[35]

The existence of digital harassment is nothing new. As far back as 1995, courts were assessing employer liability in sex discrimination cases involving email communications. In Strauss v. Microsoft Corp., the court considered a manager's and others' sexualized emails to be probative despite the fact that the statements did not relate to the adverse employment decisions at issue.[36] More tellingly, in a 2000 case out of New Jersey, the court considered sexual harassment claims emerging from digital communications on an employer-related online message board. In the case brought by a female pilot, the Court concluded that "although the electronic bulletin board may not have a physical location within [the workplace], … it should be regarded as part of the workplace. . . . [and] that if the employer had notice that co-employees were engaged on such a work-related forum in a pattern of retaliatory harassment directed at a co-employee, the employer would have a duty to remedy that harassment."[37] More recently, in a 2021 case alleging sexual

---

[29] *Id.*
[30] *Id.*
[31] Owen Hughes, *Zoom vs Microsoft Teams, Google Meet, Cisco Webex and Skype: Choosing the right video-conferencing apps for you*, TECHREPUBLIC (May 13, 2020), https://www.techrepublic.com/article/zoom-vs-microsoft-teams-google-meet-cisco-webex-and-skype-choosing-the-right-video-conferencing-apps-for-you/.
[32] *Id*.
[33] *See, e.g.* SLACK, https://slack.com/knowledge-sharing (last visited May 20, 2021) ("With a tool like Slack, your company's conversation history is at your fingertips. Instead of asking someone for information every time you need it, you can reduce repetitive questions by searching instead, saving everyone valuable time.").
[34] Sarah Gallo, *Sexual Harassment in the Remote Workplace: How Training Can Respond*, TRAINING INDUSTRY (July 21, 2020), https://trainingindustry.com/articles/compliance/sexual-harassment-in-the-remote-workplace-how-training-can-respond/.
[35] *Id.*
[36] Strauss v. Microsoft Corp., 91 Civ. 5928, 1995 U.S. Dist. LEXIS 7433, at *13 (S.D.N.Y. June 1, 1995). *See also* Owens v. Morgan Stanley & Co., 96 Civ. 9747, 1997 U.S. Dist. LEXIS 10351, at *7 (S.D.N.Y. July 16, 1997) (*aff'd*, 205 F.3d 1322 (2d Cir. 2000) (finding that, as a matter of law, a single e-mail containing racist jokes, "while entirely reprehensible, cannot form the basis for a claim of hostile work environment."); Cromwell-Gibbs v. Staybridge Suite Times Square, 16 Civ. 5169, 2017 U.S. Dist. LEXIS 95762, at *12 (S.D.N.Y. June 20, 2017) (noting that "the sending of a single offensive e-mail" does not create a hostile work environment).
[37] Blakey v. Cont'l Airlines, 751 A.2d 538, 543 (N.J. 2000).

orientation discrimination, the court permitted a "hostile work environment" claim to proceed based, in part, on an email communication in which management level employees photo shopped the plaintiff's face onto the body of a woman wearing a Mexican style dress and emailed the photo to co-workers.[38] While a number of recent cases have concluded that the legal standard for hostile work environment has not been met by the existence of a single discriminatory or harassing email, for purposes of legal liability, the courts do not distinguish between digital forms of harassment or discriminatory comments and in-person communications and conduct.[39]

Given the recent expansion in the number of employees working remotely and relying on digital communications, the EEOC's December 2020 updated guidance on the impact of COVID-19 on the workplace saw fit to remind employers that digital harassment was both possible and prohibited. "Employees may not harass other employees through, for example, emails, calls, or platforms for video or chat communication and collaboration."[40] One need only consider the Jeffrey Toobin incident, in which he was seen masturbating on a Zoom call with colleagues from The New Yorker magazine, to understand the opportunities for inappropriate online behavior in multiple forms.[41] The move to remote work and the increase in online communication did not end workplace harassment—it merely moved it to a different platform. And unlike its in-person analog, digital harassment is, more often than not, recorded, transcribed, and searchable.

B. The Tech-Infused Workplace Environment

Numerous scholars and popular media commentators have described the massive turn to technology by employers of all kinds. From hiring to worker assessment, companies are utilizing an ever-increasing number of new technological developments to assist with or take over human tasks, or to do work that humans were never capable of. A 2019 report found that "Eighty-eight percent of companies globally already use AI in some way for [Human Resources], with 100 percent of Chinese firms and 83 percent of U.S. employers relying on some form of the technology."[42]

Within HR functions, the primary use for AI is in hiring, where tech solutions include AI systems that scan and sort applicant resumes, recruit applicants from a database, and track and

---

[38] Tenorio v. Nevada, No. 2:20-cv-00517-GMN-VCF, 2021 U.S. Dist. LEXIS 20066, at *3, 19-20 (D. Nev. Feb. 2, 2021).

[39] *See*, *e.g.,* South v. Cont. Cas. Co., No. 17-cv-5741, 2018 U.S. Dist. LEXIS 167907, at *11 (S.D.N.Y. Sept. 27, 2018) (rejecting the hostile work environment claim based on a single email); Noack v. YMCA, No. 08-CV-3247, 2010 U.S. Dist. LEXIS 154053, at *35 (S.D. Tex. Mar. 1, 2010) (single e-mail does not show severity or pervasiveness of harassment that rises to the level required to support a hostile work environment claim); Lueck v. Progressive Ins., Inc., No. 09-CV-6174, 2009 U.S. Dist. LEXIS 96492, at *10 (W.D.N.Y. Oct. 19, 2009) ("The case law makes clear that the sending of a single offensive e-mail does not create a hostile work environment.")

[40] EEOC, *What You Should Know About COVID-19 and the ADA, the Rehabilitation Act, and Other EEO Laws* (Dec. 16, 2020), https://www.eeoc.gov/wysk/what-you-should-know-about-covid-19-and-ada-rehabilitation-act-and-other-eeo-laws.

[41] *See* Katherine Rosman and Jacob Bernstein, *The Undoing of Jeffrey Toobin*, N.Y. TIMES (Dec. 15, 2020) ("While working on a podcast about the presidential election for WNYC and The New Yorker with some of the magazine's other well-known journalists, including Jane Mayer and Masha Gessen, he was seen lowering and raising his computer camera, exposing and touching his penis, and motioning an air kiss to someone other than his colleagues.").

[42] Dinah Wisenberg Brin, SHRM, *Employers Embrace Artificial Intelligence for HR*, https://www.shrm.org/resourcesandtools/hr-topics/global-hr/pages/employers-embrace-artificial-intelligence-for-hr.aspx#:~:text=Eighty%2Deight%20percent%20of%20companies,some%20form%20of%20the%20technology. (discussing Mercer's Global Talent Trends 2019 report).

verify applicants in the system.[43] Companies like HiredScore, Ideal, and Eightfold, among others, emphasize the ability of their AI-driven tools to review far more applicants than human screeners could, to proactively find candidates based on prior applications, and to address hiring bias issues.[44] A 2018 survey by LinkedIn found that the appeal of AI is largely as a time-saving mechanism for reviewing hundreds or thousands of resumes and the AI's perceived ability to remove human bias from the hiring system.[45] Beyond the basic keyword searching and sorting functions, companies are increasingly turning to AI systems for their predictive capabilities. Algorithms can be built to predict worker performance based on resume length, hobbies listed, and education, with the AI learning from thousands of applicants which features reliably correlate with high performing candidates. In other words, "a machine can now identify skills and aptitudes that don't explicitly appear on a candidate's résumé."[46]

In addition to the use of AI in resume and application screening, new tech solutions are being implemented in online testing, recorded video interviews, and even video games as part of the application process. Like the machines used to predict performance based on resume features, AI is used in online testing to correlate answers to specific questions with job tenure and performance.[47] Similarly, companies ask applicants to play a video game "then use the resulting data to analyze the applicants' risk appetites, mental agility, persistence, and ability to read emotional versus contextual clues."[48] In a much-reported development, companies like HireVue are using AI and face-scanning technology to evaluate candidates from a pre-recorded video interview.[49] Despite scientific and legal critiques of this technology, it has "become so pervasive in some industries, including hospitality and finance, that universities make special efforts to train students on how to look and speak for best results. More than 100 employers now use the system, including Hilton and Unilever, and more than a million job seekers have been analyzed."[50]

Beyond hiring, AI is increasingly used to track and assess workers on the job. A 2020 report includes the following areas, in addition to talent acquisition, in which AI is useful and increasingly being implemented by HR professionals: Onboarding, Learning and Training, Cognitive-

---

[43] Bales and Stone, *supra* note 13, at 9-10. *See also* Matthew T. Bodie, Miriam A. Cherry, Marcia L. McCormick, and Jintong Tang, *The Law and Policy of People Analytics,* 88 U. COLO. L. REV. 961 (2017) (describing the use of people analytics in hiring and managing workforces and the legal and ethical dilemmas attendant to it); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017) (discussing workforce analytics and implications for anti-discrimination law).

[44] *See* HIREDSCORE, https://hiredscore.com/ (last visited Feb. 9, 2021); IDEAL, https://ideal.com/product/screening/ (last visited Feb. 9, 2021); EIGHTFOLD, https://eightfold.ai/ (last visited Feb. 9, 2021).

[45] LINKEDIN, *Global Recruiting Trends 2018*, https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/linkedin-global-recruiting-trends-2018-en-us2.pdf. Note that numerous legal scholars question the supposed unbiased nature of AI given its biased human programmers and the potentially biased sources on which it relies to build its rules. *See generally* Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671 (2020); Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519 (2018); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1 (2019).

[46] Noam Scheiber, *A.I. as Talent Scout: Unorthodox Hires, and Maybe Lower Pay*, N.Y. TIMES (Dec. 6, 2018), https://www.nytimes.com/2018/12/06/business/economy/artificial-intelligence-hiring.html.

[47] Bales and Stone, *supra* note 13, at 11.

[48] *Id.* at 13.

[49] Drew Harwell, *A face-scanning algorithm increasingly decides whether you deserve the job*, THE WASHINGTON POST (Nov. 6, 2019). *See also* Bales and Stone, *supra* note 13, at 11-12; Julie Manning Magid, *Does Your AI Discriminate?*, THE CONVERSATION (May 15, 2020), https://theconversation.com/does-your-ai-discriminate-132847.

[50] Harwell, *supra* note 49.

Supporting Decision-Making, Leadership Coaching, and Automating Administrative Tasks.[51] Using technology similar to that employed in screening resumes and predicting job performance and tenure, companies like Workday offer tech solutions that consider dozens of factors from the number of days off an employee takes to reporting structure of a unit to predict "which employees are likely to quit, which ones are likely to be disgruntled, and how the employer might retain the best employees."[52] Some companies are building their own algorithmic tools that can search through employees' social media feeds and predict, based on unrelated factors like hobbies, consumer preferences, political affiliations and the like, which employees should be given leadership opportunities, who will work best together in teams, and how costly a department will be from a health insurance perspective.[53]

In addition to predictive analytics tools that examine seemingly innocuous factors to predict performance, companies are also beginning to use sentiment analysis to assess employee moods, identify attitude shifts in a workplace, and predict inefficient periods, legal or security breaches, and other misconduct.

> AI-enabled systems will take over the task of observing and analyzing employees' mood before and after a client call. The HR can then decide if the individual needs a break or can continue. AI can also detect anxiety in a person's behavior and tone of voice. It will help the employers decide if they should look into the matter and resolve it before it is harmful to the employees and the company.[54]

Companies like KeenCorp and Teramind use AI to monitor employee communications, assessing their moods and impact on productivity, and sending tech nudges or human interventions to counteract negative attitudes, slacking, or predicted security breaches.[55] Companies are also beginning to develop facial screening technology that can detect human emotions. For example, Affectiva is "pioneering Human Perception AI: software that can detect not only human emotions, but complex cognitive states, such as drowsiness and distraction. And, in the future, it will be able to understand human activities, interactions and objects people use."[56] While the company purports to use its technology for the automotive industry and to support media analytics and biometric research, it is easy to see the applications for Human Resources. Employers could monitor employee emotional and cognitive states through their computer screens, giving nudges, mandating breaks, or even making promotion and termination decisions on the basis of this data.

As Richard Bales and Katherine Stone catalogue, in the area of electronic surveillance, there are a wide variety of new tech offerings from wristbands that track worker movements and efficiency to AI tools that analyze word patterns in digital communications.[57] There are an ever-growing number of products offering to track worker productivity. Veriato logs everything that

---

[51] *See* Khalid Durrani, *The Impact of AI in Human Resource Decision-Making Processes*, HR TECHNOLOGIST (Jan. 6, 2020), https://www.hrtechnologist.com/articles/ai-in-hr/the-impact-of-ai-in-human-resource-decisionmaking-processes/#:~:text=AI%2Dbased%20software%20can%20automate,corporate%20compliance%2C%20and%20litigation%20strategies.

[52] Bales and Stone, *supra* note 13, at 14.

[53] Leora Eisenstadt, *Data Analytics and the Erosion of the Work/Non-Work Divide*, 56 AM. BUS. L. J. 445, 472-73 (2019). *See also* Elizabeth A. Brown*, The Femtech Paradox: How Workplace Monitoring Threatens Women's Equity*, 61 JURIMETRICS J. 289, 293-295 (2021) (describing exponential growth in biometric monitoring focused on women).

[54] Durrani*, supra* note 51.

[55] Bales and Stone, *supra* note 13, at 16.

[56] AFFECTIVA, https://blog.affectiva.com/our-evolution-from-emotion-ai-to-human-perception-ai (last visited Feb. 23, 2021).

[57] *See* Bales and Stone, *supra* note 13, at 15 -21.

an employee does on his computer, tracking key strokes, time spent on websites, and application use, and can provide screenshots, real time alerts, and productivity reports.[58] "Microsoft's MyAnalytics amalgamates data from a worker's emails, calendars, and phones to calculate how the worker spends her time, how often she is in touch with key contacts, and whether she multitasks too frequently."[59] Amazon has patented a "haptic wristband" that tracks employee movements and can nudge the employee to work faster or more efficiently.[60]

While it may be tempting to view these technological developments as scattered offerings and not a sign of significant industry changes, several recent surveys demonstrate the widespread enthusiasm for AI in the Human Resources field in particular. A 2017 study by IBM made this statement:

> Our study reveals that the market for cognitive solutions in HR is set to increase notably over the next three years: Sixty-six percent of CEOs believe cognitive computing can drive significant value in HR, and almost 40 percent expect their HR function to adopt cognitive solutions during that time. Business leaders understand that cognitive computing is a critical differentiator in the ongoing war for talent.[61]

In talent development specifically, the IBM study suggested that cognitive computing would be particularly useful in assisting HR professionals to "understand employee sentiment and more rapidly identify emerging issues."[62] Because of technology's ability to monitor numerous data sources including internal digital communications, external social media platforms, computer and internet use, and employee facial expressions, AI can "search for potential issues or employee concerns", serving as "canaries in a virtual coal mine," identifying hot topics and longer-term trends that could affect employee morale and performance."[63]

While the IBM study demonstrates leadership's openness to incorporating AI in HR functions, a recent Oracle study suggests that employees are equally enthusiastic. Titled "From Fear to Enthusiasm," Oracle's 2019 study surveyed over "8,000 HR leaders, managers, and other employees across 10 countries on their attitudes toward and behaviors regarding AI."[64] Like the IBM study, Oracle's survey demonstrated that HR leaders are most optimistic about using AI.[65] More than that, the study also showed that "as many as 50% of . . . survey respondents this year said they're currently using some form of AI at work. That's an impressive jump from the 32% who said this in last year's survey."[66] Perhaps more surprising is the finding that employees are embracing the technology as well.

---

[58] *See* VERIATO, https://www.veriato.com/products/veriato-vision-employee-monitoring-software (last visited May 18, 2021).
[59] Bales and Stone, *supra* note 13, at 16-17.
[60] *Id.* at 17. *See also* Ceylan Yeginsu, *If Workers Slack Off, the Wristband Will Know. (And Amazon Has a Patent for It.)*, N.Y. TIMES (Feb. 1, 2018), https://www.nytimes.com/2018/02/01/technology/amazon-wristband-tracking-privacy.html
[61] IBM INSTITUTE FOR BUSINESS VALUE, EXTENDING EXPERTISE - HOW COGNITIVE COMPUTING IS TRANSFORMING HR AND THE EMPLOYEE EXPERIENCE 5 (2017), https://www.ibm.com/downloads/cas/QVPR1K7D.
[62] *Id* at 13.
[63] *Id.*
[64] ORACLE & FUTURE WORKPLACE, FROM FEAR TO ENTHUSIASM – ARTIFICIAL INTELLIGENCE IS WINNING MORE HEARTS AND MINDS IN THE WORKPLACE 2 (2019), https://www.oracle.com/a/ocom/docs/applications/hcm/ai-at-work-ebook.pdf.
[65] *Id.* at 4.
[66] *Id.* at 2.

> Workers apparently now trust robots more than they trust their managers. Consider this: As many as 82% of our survey respondents said they think robots can do certain types of work better than their managers. A whopping 64% said they'd trust a robot more than their manager, and 50% have turned to a robot instead of their manager for advice, with nearly 25% saying they "always" or "very often" ask AI questions rather than over asking their boss.[67]

In response to a question about how their companies are already using AI, the survey showed the following: 31% are using it to collect data on employees and customers, 28% are using it to develop software for training, 24% use AI to manage customer-support replies, 22% use it to operate digital assistants, 21% use it to predict hiring success rate and employee retention, and 17% use it to process job applications.[68] It is abundantly clear that the adoption of AI in HR functions is not a passing trend but rather an increasingly permanent fixture of the industry.

It is in the context of these dual realities— the exponential increase in digital communications and the overwhelming embrace of AI in hiring and managing employees—that companies have now proposed the use of AI to detect and report sexual harassment of all kinds. Like the existing critiques of AI-based technology in the workplace, this move into AI-driven harassment detection brings ethical and legal problems that must be explored. Unlike prior critiques that have often focused on the biases that may be inherent in AI innovations, the development of #MeTooBots necessitates an examination of the role of human reporters in both employer defenses and employee protections and the implications of replacing that human reporter with an automated bot.

II. #MeTooBots' Capabilities: the Why and the How

In order to understand the potentially dramatic impact of AI harassment monitors on the workplace and legal doctrine, it is necessary to first examine the way in which AI functions in this space. The technology behind #MeTooBots has been hailed by some as part of a promising future and by others as a woefully inadequate replacement for human interventions. This section will explore what companies attempt to do with this technology and the ways in which it can and cannot mimic or improve upon human abilities.

A. Why create #MeTooBots?

The development of #MeTooBots seems to have been a response to the acknowledgment of digital harassment as a workplace reality and an attempt to solve two major issues for businesses—increased attention and pressure on employers to rectify "toxic workplaces" and dramatic underreporting of workplace harassment. Notwithstanding the fact that digital and online harassment, as described above, has been acknowledged by Human Resources professionals, the EEOC, and the popular media, some companies may operate on the mistaken notion that the move to remote work will reduce harassment in the workplace overall.[69] Born of a faulty understanding of harassment as being solely physical or somehow requiring face-to-face interactions, the reality

[67] *Id.* at 10.
[68] *Id.* at 13.
[69] *See supra* text accompanying notes 34-41.

is that harassment encompasses a wide range of communications, behaviors, and interactions that includes digital, verbal, and physical interactions.[70] In fact, recent reports suggest that people go beyond harassing their human co-workers in online platforms and will even harass digital assistants and bots. Apple's Siri and Microsoft's Cortana have both been on the receiving end of verbal abuse. CNN reports, "[a] side effect of creating friendly female personalities is that people also want to talk dirty, confess their love, role play, or bombard them with insults."[71] Given this tendency, the fact that workers would behave inappropriately with co-workers online as they do in person is a fairly obvious notion, particularly since many workers do not think about the fact that their online behavior is being recorded, transcribed, or observed.[72]

In addition to the reality of online or digital harassment, the first major problem to which the creators of AI harassment monitors are responding is the #MeToo Movement's increase in attention to harassment and the concomitant pressure on companies to rectify the problem of harassment in the workplace by "catching" and eliminating harassers. The phrase "Me too," coined by activist Tarana Burke in 2006 took on new life in 2017 when numerous actresses accused producer Harvey Weinstein of sexual harassment, followed by accusations against and resignations or other career consequences for Amazon Studios head Roy Price, Fox News host Bill O'Reilly and chairman and CEO Roger Ailes, actor Kevin Spacey, comedian Louis C.K., public radio personality Garrison Keillor, U.S. Senator Al Franken, celebrity chef Mario Batali, and many other figures in business, media, sports, education, government, entertainment, and politics.[73] The movement led to the unearthing of rampant sexual harassment across industries and put significant pressures on workplaces of all kinds to improve in this area.[74] This pressure is both external and internal. Companies are responding to the brand reputation issues associated with the uncovering of harassment in a workplace and the increased costs of investigations and litigation.[75] They are also beginning to recognize the impact of harassment on employee health and well-being, affecting

[70] *See* EQUAL EMPLOYMENT OPPORTUNITY COMMISSION (EEOC), HARASSMENT, https://www.eeoc.gov/harassment ("Offensive conduct may include, but is not limited to, offensive jokes, slurs, epithets or name calling, physical assaults or threats, intimidation, ridicule or mockery, insults or put-downs, offensive objects or pictures, and interference with work performance.")

[71] Omer Tene and Jules Polonetsky, *Taming The Golem: Challenges Of Ethical Algorithmic Decision-Making*, 19 N.C. J.L. & TECH. 125, 153 (2017) (quoting Heather Kelly, Even Virtual Assistants Are Sexually Harassed, CNN (Feb. 5, 2016, 10:41 AM), http://money.cnn.com/2016/02/05/technology/virtual-assistants-sexual-harassment/index.html.).

[72] *See* Eisenstadt, *Data Analytics, supra* note 53, at 469 (describing the reality that despite the fact that workers know their computers and phones are owned by employers and subject to monitoring, "this technical knowledge rarely stops employees from using those devices for personal e-mail, personal social media networking, and generally communicating thoughts, pictures, and other information that they would not actively want to share with their employers.").

[73] *See* Chicago Tribune, *#MeToo: A timeline of events*, CHICAGO TRIBUNE (Feb. 4, 2021), https://www.chicagotribune.com/lifestyles/ct-me-too-timeline-20171208-htmlstory.html.

[74] *See* Riley Griffin, Hannah Recht and Jeff Green, *#MeToo: One Year Later*, BLOOMBERG (Oct. 5, 2018), https://www.bloomberg.com/graphics/2018-me-too-anniversary/.

[75] *See, e.g.* Mike Isaac, *Uber Investigating Sexual Harassment Claims by Ex-Employee*, N.Y. TIMES (Feb. 19, 2017), https://www.nytimes.com/2017/02/19/business/uber-sexual-harassment-investigation.html; Steven Overly, *Uber hires Eric Holder to investigate sexual harassment claims*, WASH. POST (Feb. 21, 2017); Olivia Solon, *Uber fires more than 20 employees after sexual harassment investigation*, THE GUARDIAN (June 6, 2017), https://www.theguardian.com/technology/2017/jun/06/uber-fires-employees-sexual-harassment-investigation.

absences from work and insurance costs.[76] Harassment can also impact productivity since "[h]arassed people [are] less satisfied with their jobs, and [are] more likely to want to leave."[77]

The results of this attention have been demands for increased transparency, diversity, and most profoundly, a focus on company culture.

> More board members will ask tough questions about efforts to increase diversity and programs to foster a respectful culture. . . . We'll see similar shifts in funding, with more venture capitalists prioritizing company culture and diversity in investment decisions. Soon it will no longer be uncommon for a VC to ask about diversity at the top or plans to create a more inclusive workplace. While not every investor will focus on culture and diversity, the smart ones will (and some already are).[78]

Most relevant to this Article, the #MeToo Movement increased the value of data analytics in efforts to detect and fix problems before they take over. As one commentator put it, "Rather than solving a culture problem after the fact, you can identify issues early on and course correct. . . . Whether it's adopting HR tech for your organization, asking difficult questions, or focusing on inclusion initiatives, each of us has a role in creating a strong company culture."[79]

At the same time, creators of AI harassment monitors are responding to the unfortunate reality that employees across industries often fail to report harassment. According to a meta-analysis of studies, "only a quarter to a third of people who have been harassed at work report it to a supervisor or union representative, and 2 percent to 13 percent file a formal complaint."[80] Researchers have identified numerous barriers to reporting including unclear or confusing processes, reluctance to discuss personal issues or sexual behavior, and most commonly, lack of trust and fear of retaliation, "citing negative consequences when others reported incidents."[81] A N.Y. Times investigation at Fox News showed that "women who worked at Fox said they didn't complain to human resources because they feared they would be fired."[82] A recent survey at the FDIC suggested significant underreporting of harassment and concluded that fear of retaliation was one of the primary reasons for the underreporting.[83] Similar findings emerged from an investigation at the State Department, which "revealed 47 percent of department employees surveyed who experienced or observed harassment failed to tell the State Department's internal bureaus that handle misconduct complaints."[84] And again, the reasons for the underreporting include "lack of confidence in the

---

[76] Julia Shaw and Camilla Elphick, In the #MeToo Era, a chatbot can help people report workplace harassment, THE CONVERSATION (March 19, 2018), https://theconversation.com/in-the-metoo-era-a-chatbot-can-help-people-report-workplace-harassment-92565.

[77] *Id.*

[78] Spencer Rascoff, *3 Ways #MeToo Will Influence the Business World in 2018*, INC. (Jan. 5, 2018), https://www.inc.com/spencer-rascoff/3-ways-metoo-will-influence-business-world-in-2018.html.

[79] *Id.*

[80] Claire Cain Miller, *It's Not Just Fox: Why Women Don't Report Sexual Harassment*, N.Y. TIMES (April 10, 2017) (referencing Lilia M. Cortina and Jennifer L. Berdahl, *Sexual Harassment in Organizations: A Decade of Research in Review* in THE SAGE HANDBOOK OF ORGANIZATIONAL BEHAVIOR (2008), https://lsa.umich.edu/psych/lilia-cortina-lab/Cortina&Berdahl.2008.pdf).

[81] Shaw and Elphick, *supra* note 76.

[82] Miller, *supra* note 80.

[83] *See* Federal Manager's Daily Report, *Harassment, Fear of Retaliation Reported at FDIC*, FEDWEEK (July 17, 2020), https://www.fedweek.com/federal-managers-daily-report/harassment-fear-of-retaliation-reported-at-fdic/.

[84] Pranshu Verma, *Sexual Harassment Underreported at State Department, Report Says*, N.Y TIMES (Oct. 2, 2020), https://www.nytimes.com/2020/10/02/us/politics/sexual-harassment-underreported-state-department.html.

department's ability to resolve complaints, fear of retaliation and reluctance to discuss the harassment with others."

This fear of retaliation is decidedly justified. In 2020, "Workers alleging employers unlawfully retaliated against them once again topped the charts of claims filed with the EEOC . . . , a trend that's held strong for at least three years, and now through a pandemic."[85] At least one court has recognized this fear of retaliation and considered it a reasonable justification for an employee to fail to report her harassment with potential impacts on employer liability. In *Minarsky v. Susquehanna County*,[86] the Third Circuit acknowledged the #MeToo stories that have shared this fear and noted that "a jury could conclude that [an] employee's nonreporting was understandable, perhaps even reasonable."[87]

Given the realities of digital harassment, massive underreporting by victims of harassment, and increasing pressure to take action in this area, it is perhaps unsurprising that companies are looking for "easy" solutions. It is in this context that AI-based harassment monitors have been developed and marketed—the #MeTooBots can scour every digital communication and report anything problematic to HR or in-house counsel, thereby eliminating the need for a human reporter and alleviating the problem of underreporting while taking action to "fix" toxic workplaces or remove problem workers. But this approach is laden with problems both technical and ethical/legal.

B. What do #MeTooBots Do?

To understand what #MeTooBots do and how they do it, it is essential to understand the process of algorithmic decision-making and machine learning. In general, AI is considered to be an approach that uses "technology to automate tasks that 'normally require human intelligence.' . . . [T]he technology is often focused upon automating specific types of tasks: those that are thought to involve intelligence when people perform them."[88]

Among the companies that offer AI-based harassment monitors are NexLP and Aware.[89] As Aware explains, the use of developing technology to monitor and secure data has been common place for decades. What is new is the use of AI to enhance the capabilities of this technology. "An AI algorithm uses Machine Learning (ML) to adapt over time, making it more 'smart' and 'secure' over time."[90] Applying this technology to the human resources function, Aware suggests a dramatic shift in the accuracy and breadth of its detection capabilities.

> Without artificial intelligence algorithms, monitoring platforms on the market leverage only keyword matching to identify suspicious instances of digital communications. . . . By leveraging AI-infused technology, such as Aware,

[85] Paige Smith, *EEOC Again Sees Prevalence of Workplace Retaliation Claims*, BLOOMBERG LAW (Feb. 26, 2021).
[86] 895 F.3d 303 (3rd Cir. 2018).
[87] *Id.* at 314 (The court found that "[i]f a plaintiff's genuinely held, subjective belief of potential retaliation from reporting her harassment appears to be well-founded and a jury could find that this belief is objectively reasonable … the court should leave the issue for the jury to determine at trial.").
[88] Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U.L. REV. 1305, 1307 (2019).
[89] *See infra* note 90, 93. *See also* Sejuti Das, *AI To Combat Sexual Harassment With Chatbots, Apps & Trained Algorithms*, ANALYTICS INDIA MAGAZINE (June 1, 2020), https://analyticsindiamag.com/ai-to-combat-sexual-harassment-with-chatbots-apps-trained-algorithms/.
[90] AWARE HQ, USING AI TO IDENTIFY AND REDUCE SEXUAL HARASSMENT AT WORK (May 2, 2018), www.awarehq.com/blog/identifying-and-reducing-workplace-sexual-harassment-with-ai.

> organizations can identify, investigate, and handle offensive communications in the early stage – without requiring the victim to report the incident to a superior.[91]

NexLP, recently acquired by Reveal,[92] offers a product it calls Story Engine™ I³, which can "actively monitor enterprise communications to turn disparate data into decisive action."[93] The company offers multiple "use cases" for the technology, suggesting its application in the compliance and HR spaces.[94] It specifically proposes its use to detect and report issues involving "discrimination, corporate social responsibility, loyalty risk index & retention, [and] sexual harassment prevention."[95] Like Aware, NexLP's AI-based products aim to arm employers with tools to make early detection and prevention easier and more cost-effective. As NexLP's promotional materials explain:

> In today's volatile world employee and customer demeanor can change at the slightest indication of scandals, attacks, and simply rumor. Wouldn't it be great to visualize employee sentiment, what topics are trending in your organization, and where the biggest changes are occurring in real time? NexLP's off-the-shelf cognitive analytics platform uses next generation artificial intelligence and machine learning to detect signals at their earliest stage—giving you time to react.[96]

While the companies that advertise these bots do not, for obvious reasons, describe exactly how the technology works, we can make some basic assumptions based on other AI-based monitoring technology. In general, the AI-based technologies are not actually "thinking" but rather producing "intelligent results without intelligence" by "detecting patterns in data and using knowledge, rules, and information that have been specifically encoded by people into forms that can be processed by computers."[97]

The form of AI technology that is most likely employed by the harassment monitors and the broad category of techniques that is most often referred to in monitoring technology is machine learning. "Most machine-learning methods work by detecting useful patterns in large amounts of data. These systems can then apply these patterns in various tasks, such as driving a car or

---

[91] *Id.*

[92] Jennifer Fournier, *Reveal Acquires NexLP to become the leading AI-powered eDiscovery Solution*, CISION PR NEWSWIRE (Aug. 11, 2020), https://www.prnewswire.com/in/news-releases/reveal-acquires-nexlp-to-become-the-leading-ai-powered-ediscovery-solution-844046460.html.

[93] NEXLP, ARTIFICIAL INTELLIGENCE SOLUTIONS - DYNAMIC OFF THE SHELF PRODUCTS, https://www.nexlp.com/products.

[94] *Id.*

[95] *Id.*

[96] *Id.* NexLP produces "out of the box" solutions to detect workplace problems. "These new AI models are designed to uncover different aspects of employment-related behaviors, across any industry. To help our clients continue to achieve rapid contextual data review, our Data Science team is working hard to publish new models on a regular basis. Our newly available models are:

> Employment & Career Advancement: Identifies conversations about promotions, job opportunities, and workplace performance.
>
> Comments on Appearance: Identifies conversations related to a person's attire or physical attributes. These conversations may have positive or negative connotations, and may exist in the form of jokes, compliments, or advice.
>
> Sexually Explicit Comments: Identifies conversations that include descriptive language related to sexual acts or inappropriate behavior. These conversations may be general in nature, or targeted at a specific person or situation.

Ye Chen, *NexLP Releases 3 New AI Models for Faster Discovery* (April 24, 2020), https://www.nexlp.com/blog/story-engine-models-faster-discovery.

[97] Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1308 (2019).

detecting fraud, in ways that often produce useful, intelligent-seeming results."[98] One example that is particularly useful in understanding harassment monitors is to consider commonly-used e-mail spam filters. Harry Surden explains:

> Most e-mail software uses machine learning to automatically detect incoming spam e-mails (i.e. unwanted, unsolicited commercial e-mails) and divert them into a separate spam folder. How does such a machine-learning system automatically identify spam? Often the key is to "train" the system by giving it multiple examples of spam e-mails and multiple examples of "wanted" e-mails. The machine-learning software can then detect patterns across these example e-mails that it can later use to determine the likelihood that a new incoming e-mail is either spam or wanted. For instance, when a new e-mail arrives, users are usually given the option to mark the e-mail as spam or not. Every time users mark an e-mail as spam, they are providing a training example for the system. This signals to the machine-learning software that this is a human-verified example of a spam e-mail that it should analyze for telltale patterns that might distinguish it from wanted e-mails. . . . One common approach simply uses word probabilities. In that technique, the system attempts to detect words and phrases that are more likely than average to appear in a spam e-mail. . . .We can think of this as an intelligent result because this is roughly what a person would have done had he quickly scanned the e-mail, noticed [the words commonly found in spam emails] and decided it was spam.[99]

In essence, machine learning programs are provided with an initial set of rules and then permitted to program themselves as more and more data is presented. They are constantly improving over time as they further hone the patterns from larger and larger bodies of data. [100]

As Surden points out, it is important to understand that many AI systems, including #MeTooBots as advertised, include some human decision-making.[101] "This system design is known as having 'a human in the loop.'"[102] The AI system will function autonomously up to a point where it is trained to involve a human in the process.[103] In the context of the AI-based harassment monitors, the AI appears to function autonomously until it identifies inappropriate digital communication, at which point, it does not dole out punishment on its own but rather reports the occurrence to Human Resources or in-house counsel for review.[104]

Lastly, the AI-based harassment monitors likely depend also on "natural language processing" (NLP) in order to make sense of the words, phrases and patterns that they detect. "NLP is the study of computational linguistics, which includes natural language understanding (NLU) and natural language generation (NLG)."[105] Whereas machine learning describes the process of

---

[98] *Id.* at 1311.
[99] *Id.* at 1312-13.
[100] *Id.* at 1314. In addition to machine learning, some AI tools use what Surden refers to as "logical rules and knowledge representation" or a hybrid approach that combines both types of AI. *Id.* at 1316-19. "Knowledge representation" systems typically involve programmers providing the computer with a series of rules "that represent the underlying logic and knowledge of whatever activity the programmers are trying to model and automate." *Id.* at 1316. In hybrid form, the AI uses a combination of machine learning to track and process patterns in data with a series of rules and knowledge translated into computer-processable form. *Id.* at 1320.
[101] *Id.* at 1320.
[102] *Id.*
[103] *Id.*
[104] *See* Woodford, *supra* note 1.
[105] Brian S. Haney, *Applied Natural Language Processing For Law Practice*, 2020 B.C. INTELL. PROP. & TECH. F. 1, 4 (2020).

pattern recognition and prediction based on those patterns, NLP refers to the computer's ability to "process, understand, and generate language representations as well as humans."[106] Because human language makes meaning based on more than words alone, NLP involves teaching computers to understand context, syntax, and semantics.[107] The goal of NLP tools is to allow the computer to parse text and predict its meaning – whether it expresses positive or negative sentiment, whether it is inappropriate or harmless, and so on.[108] Regardless of the specific tool the AI relies upon or, more likely, the combination of tools, this technology allows companies to monitor, flag, and automatically report inappropriate communications without involving the sender or receiver in that process.[109]

III. Problems with #MeToo Bots: Functional, Legal, and Ethical

As described above, #MeToo Bots that use AI-based technologies to monitor, detect, and report digital harassment emerged from a need to take action on the massive problem of unreported sexual harassment in the workplace. Despite its seemingly good intentions, this technology has numerous technical inadequacies and more importantly for this article, creates several legal and ethical problems. This Section details both the functional and legal/ethical obstacles and discusses existing scholarship that brings to light similar concerns with related uses of AI in the workplace.

A. Functional Problems with AI-based Harassment Monitors

The primary focus of this article is the legal/ethical problems inherent in using AI-based harassment monitors. Nonetheless, it is important to understand the ways in which the technology itself is still limited in its ability to perform its designated function both as a general matter and because those limitations, in fact, impact both the legal rights of employees on whom it is used and the legal liability of employers who use it. This is particularly true for already marginalized employees who may find their communications targeted by the AI because of the functional problems inherent in these tools.

---

[106] *Id.*

[107] *Id.*

[108] NATASHA DUARTE, EMMA LLANSO, AND ANNA LOUP, MIXED MESSAGES? THE LIMITS OF AUTOMATED SOCIAL MEDIA CONTENT ANALYSIS 9, CENTER FOR DEMOCRACY & TECHNOLOGY (Nov. 2017), https://cdt.org/files/2017/11/Mixed-Messages-Paper.pdf.

[109] It is important to also understand that although referred to as "#MeTooBots" these AI-based harassment monitors are distinct from chat bots that may also use some form of AI-based technology. These chat bots allow a victim to report an incident to a robot instead of another person, creating an increased comfort level and sense of confidentiality. " The Spot chatbot – designed to improve the reporting process for those who have experienced or witnessed workplace harassment or discrimination – guides the participant through an evidence-based cognitive interview without requiring them to talk to a human." *Bots Supporting the #MeToo Movement*, INSTABOT BLOG (Sept. 28, 2018), https://blog.instabot.io/instabot-blog/2018/9/bots-supporting-the-metoo-movement. Scientists in Nigeria developed a chatbot named Biri that allows victims of physical and sexual abuse to record their stories before offering a connection to a human counselor. *Id.* Likewise, an Indian company created Me2Bot, which provides a confidential platform for victims of workplace harassment to report their experiences. The bot then sends users resources to help them deal with the particular incident. *Id.* These tools offer an AI-based platform with which victims can communicate, rather than an algorithm that searches through digital communications to detect harassment. For chat bots to function, victims need to come forward and report incidents of harassment or assault. For the #MeTooBots that function as harassment monitors to function, no such action is required; the victim can remain passive throughout the process.

1. Nuanced Human Communication

The most essential deficiency in #MeTooBots' functionality emerges from a basic problem that arises when AI is tasked with "understanding" language—computers are relatively bad at comprehending the nuances of human communication. The meaning of verbal communications is derived from contextual clues including the social and cultural identity of the speakers, the nature of the communication, the relationship between the parties, the history of such communications, and potentially hundreds of other factors.[110] However, "AI can only reliably conduct basic story analysis, meaning it is taught to look for specific triggers. It cannot go beyond that parameter and cannot pick up on broader cultural or unique interpersonal dynamics."[111] This problem arises frequently in the context of algorithms used to screen job applicants' social media profiles. Such tools "have limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the speaker. Even the most advanced technology companies struggle to define and automatically identify 'toxic' content."[112]

This problem is compounded in the arena of sexual harassment because, as Brian Subirana points out in *The Guardian*, communication of a sexual nature is particularly complex. "There's a type of harassment that is very subtle and very hard to pick up. We have these training courses [about harassment] . . . and it requires the type of understanding that AI is not yet capable of."[113] The problem is also amplified in the harassment context because the line between acceptable and inappropriate communication is not always entirely clear to the people involved. That line may be dependent on context and on the relationship between and prior history of the parties involved.[114]

This is particularly true when legal parameters are grafted onto this already fraught area. Two judges considering nearly identical verbal communications may classify them differently with respect to the law. Where one judge may find a particularly egregious sexual comment to be unlawful harassment, another may view it as a solitary event that is not "severe or pervasive"

---

[110] *See generally* Leora F. Eisenstadt, *The N-Word at Work*: *Contextualizing Language in the Workplace*, 33 BERKELEY J. EMP. & LAB. L. 299 (2012) (discussing contextual nature of language and its meaning).

[111] Woodford, *supra* note 1.

[112] MIRANDA BOGEN & AARON RIEKE, UPTURN, HELP WANTED: AN EXAMINATION OF HIRING ALGORITHMS, EQUITY, AND BIAS 40 (Dec. 2018), https://www.upturn.org/reports/2018/hiring-algorithms [https://perma.cc/277F-ZG97]. In fact, among studies that assess the accuracy of natural language processing tools in determining the meaning of text, "the highest accuracy rates reported hover around 80%, with most of the high-performing tools achieving 70 to 75% accuracy." Duarte, *et. al.*, *supra* note 108, at 5. While this is remarkable from a technological standpoint, it means that a significant number of determinations made by the AI is "wrong," and depending on the particular use of the AI, this inaccuracy rate can have unacceptable consequences for those individuals whose communications are improperly judged. *Id.*

[113] Woodford, *supra* note 1.

[114] *See* Eisenstadt, *N-Word at Work*, supra note 110, at 316-320. *See also* DEBORAH TANNEN, GENDER & DISCOURSE 19-20 (1994) ("In analyzing discourse, many researchers operate on the unstated assumption that all speakers proceed along similar lines of interpretation, so a particular example of discourse can be taken to represent how discourse works for all speakers. For some aspects of discourse, this is undoubtedly true. Yet a large body of sociolinguistic literature makes clear that, for many aspects of discourse, this is so only to the extent that cultural background is shared. . . . Thus, a strategy that seems, or is, intended to dominate may in another context or in the mouth of another speaker be intended or used to establish connection. Similarly, a strategy that seems, or is, intended to create connection can in another context or in the mouth of another speaker be intended or used to establish dominance.").

enough to be considered unlawful under the relevant legal doctrine.[115] Adding AI to this mix yields problematic results. Natural language processing tools, the primary technology used by AI-based monitors, require clear rules that guide the system on which words, phrases, and combinations to flag.

> [T]oday's AI produces results by "detecting patterns in data and using knowledge, rules, and information that have been specifically encoded by people into forms that can be processed by computers." It "excels in narrow, limited settings, like chess, that have particular characteristics--often where there are clear right or wrong answers, where there are discernable underlying patterns and structures, and where fast search and computation provides advantages over human cognition."[116]

In contrast to a game of chess, application of anti-discrimination law to instances of sexual harassment lack precise rules on which the AI can rely.[117] Attorneys and courts make judgments in cases based on experience and common sense, judging the communication by its words and context.[118] The system relies on evaluators' ability to be flexible in their understanding and judgments. They are "sensitive to context, both to extenuating circumstances in individual cases and shifts in social norms over time, and can flexibly apply legal rules."[119] As Rebecca Crootof explains, AI-based tools are "brittle," making them ill-equipped to comprehend and judge nuanced, contextual, and subtle forms of human communication.[120] The notion that AI can competently identify inappropriate or unlawful communications in the sexual harassment sphere ignores this basic functionality problem.

2. Signal and Bias Problems

In addition to these basic problems is the related issue that Natural Language Processing tools, despite often being relied upon to counteract human implicit bias, may often, in fact, "amplify social bias reflected in language."[121] The AI tools learn the "rules of interpretation" from the data provided by their human coders. If the data provided is biased in any way, the results produced

---

[115] *See* Jeffrey R. Boles, Leora Eisenstadt, and Jennifer M. Pacella, *Whistleblowing in the Compliance Era*, 55 GEORGIA L. REV. 147, 165 (2020) (describing the varied ways that courts treat similar facts when applying harassment and antidiscrimination doctrine).

[116] Rebecca Crootof**, "***Cyborg Justice" and the Risk of Technological-Legal Lock-In*, 119 COLUM. L. REV. FORUM 233, 237-38 (Nov. 20, 2019) (quoting Surden, *supra* note 97, at 1308 (noting that "[t]his description of AI programming roughly encompasses both machine learning and rule-based systems, as most AI systems today incorporate elements of both.").

[117] Woodford, *supra* note 1. *See also* Duarte et al., *supra* note 108, at 16.

[118] *See* Oncale v. Sundowner Offshore Servs., 523 U.S. 75, 81 (1998) ("We have emphasized . . . that the objective severity of harassment should be judged from the perspective of a reasonable person in the plaintiff's position, considering 'all the circumstances.' In same-sex (as in all) harassment cases, that inquiry requires careful consideration of the social context in which particular behavior occurs and is experienced by its target.")

[119] Crootof, *supra* note 116, at 238.

[120] *Id*. at 239-40 (describing the difficulty of using AI to automate enforcement of even simple laws and pointing to unsuccessful attempts to automate speeding laws). The NLP tools that are marketed as "off the shelf" are particularly vulnerable to this critique of inflexibility. "Language use can vary considerably across and within social media platforms, demographic groups, and topics of conversation. The language people use in captions when sharing images of their pets on Instagram has very different characteristics from the language used to discuss major geopolitical events on Facebook. A tool trained to recognize the former cannot be reliably applied to analyze the latter. NLP tools must be trained to recognize the particular type (or "domain") of speech they will be used to analyze; otherwise their performance will suffer." Duarte et al., *supra* note 108, at 4.

[121] Duarte et al., *supra* note 108, at 4.

by the AI will be biased as well.[122] "Several studies have found, for example, that machine learning models reflect or amplify gender bias in the text used to train them. This type of bias could lead to content moderation decisions that disproportionately censor or misinterpret the speech of certain groups, such as marginalized groups or those with minority views."[123]

The "signal problems" typically arise when the data provided to the machine learning tool does not sufficiently include speech patterns of underrepresented groups.[124] Unless the AI tool is trained on speech patterns encompassing diverse races, ethnicities, genders, and socioeconomic groups, harassing communications will be missed, and innocuous comments will be mischaracterized as inappropriate or unlawful harassment. For example, ProPublica did a study in which it created a tool using word2vec, which creates "word embeddings" based on how words are related to each other and the context and order in which they appear, and uses the resulting "word maps" to make meaning.[125] ProPublica's tool was trained on a variety of "media diets" to test the output based on the training data.

> For the term "imma," frequently used in African- American Vernacular English, only the algorithm trained on a digital media diet recognized the word and produced results. The outputs for "imma" were mostly offensive words that would likely be associated with hate speech or threats, even though "imma" simply means "I'm going to."[126]

Were an AI-based harassment monitor to be trained on similar data, it might identify this and other words from African-American Vernacular English (AAVE) as threatening or harassing despite the prevalence of completely innocuous uses. Similarly, when NLP tools are imposed on non-English text or text in languages that are not well-represented in online sources, the accuracy rate again suffers because of the limited data on which they are trained.

> A major limitation of NLP today is the fact that most NLP resources and systems are available only for high-resource languages (HRLs) such as English, French, Spanish, German, and Chinese. In contrast, many low-resource languages (LRLs)—such as Bengali, Indonesian, Punjabi, Cebuano, and Swahili—spoken and written by millions of people have no such resources or systems available.[127]

Many NLP tools are trained to avoid non-English text altogether.[128] This is particularly problematic when the NLP tool misidentifies dialects as being non-English. For example, popular NLP tools tend to misidentify African American Vernacular English (AAVE) as non-English.

---

[122] *See* Ajunwa, *supra* note 45, at 1685-86.

[123] Duarte et al., *supra* note 108, at 4. *See also* Cade Metz, *Using A.I. to Find Bias in A.I.*, N.Y. TIMES June 30, 2021 (describing the problem of bias in artificial intelligence which is "now facing increasing scrutiny from regulators and is a growing business for start-ups and tech stalwarts").

[124] *See* Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013, 12:40 AM), https://foreignpolicy.com/2013/05/10/think-again-big-data [https://perma.cc/T4F4-V54J] (cited in Ajunwa, *supra* note 45, at 1686). *See also* Dirk Hovy & L. Shannon Spruit, *The Social Impact of Natural Language Processing*, PROC. ASSOC. FOR COMPUTATIONAL LINGUISTICS (2016); Maider Lehr, Kyle Gorman, & Izhak Shafran, *Discriminative Pronunciation Modeling for Dialectal Speech Recognition*, PROC. INTERSPEECH (2014).

[125] Duarte et al., *supra* note 108, at 11.

[126] Duarte et al., *supra* note 108, at 14. Interestingly, this same word also means "mother" or "mom" in Hebrew and is sometimes used by otherwise English-speaking Jews to refer to their mothers (like this author's children), demonstrating again the complexity of human written and verbal communication.

[127] Julia Hirschberg & Christopher D. Manning, *Advances in Natural Language Processing*, 349 SCIENCE 261, 261 (July 17, 2015), https://cs224d.stanford.edu/papers/advances.pdf. "High-resource languages" refers to the quantity of training data available in those languages. *See* Duarte et al., *supra* note 108, at 14.

[128] Duarte et al., *supra* note 108, at 15.

"One system identified examples of AAVE as Danish with 99.9% confidence".[129] If the NLP tool is trained to skip over non-English text, significant portions of dialect-based communications may be overlooked altogether. When AI-based tools attempt to transcribe spoken communications (as Zoom and other video conferencing tools provide), the ability of the tools to "understand" dialects and gender and racial speech patterns is again problematic. "A 2017 study found that YouTube auto-captioning had a higher error rate for captioning female speakers than for male speakers in videos."[130]

These functional and bias-amplification problems impact the accuracy of AI-based tools' ability to "understand" or make meaning out of human communication whether spoken or written. The failures mean that in numerous instances, an AI-based monitor will improperly investigate and report benign communications, a danger that will likely be borne more frequently by employees of color and any groups whose communication is underrepresented in the training process. This, in turn, exposes already marginalized employees to additional discrimination and harassment in the form of unnecessary, disruptive, and demeaning investigations. The workplace may understandably feel even more hostile to employees whose communications are regularly flagged and investigated even if they are ultimately exonerated. And whereas under normal circumstances, a human resources officer might quickly discount such a benign communication, when AI is involved, this instinct is often overridden by what Ifeoma Ajunwa refers to as the "unquestioning belief in data objectivity."[131] Thinking that if the AI suggested this was problematic, it is at least deserving of further investigation. Those who rely on these algorithms begin to trust them implicitly, becoming blind to the ways in which AI is incapable of understanding nuanced language and the ways in which it fails to accurately capture the breadth of human communication.

B. Legal and Ethical Problems Inherent in #MeTooBots

When evaluating new tools (often involving new technology) that employers deploy in the workplace, there is a tendency in popular media and legal scholarship to focus primarily on the impact on employee rights and protections. That exclusives focus is a mistake in this and other areas because a closer look at the technology exposes problematic implications for employer and employee alike. Employers' exposure to greater liability and employees' increased vulnerability to lawful retaliation, in this case, creates an alignment of interests that, if understood, cautions against blindly adopting #MeTooBot technology.

1. Impact on Employer Liability

Although sexual harassment is prohibited by Title VII of the Civil Rights Act of 1964, it was not until 1986 that the Supreme Court defined sexual harassment as a form of unlawful sex discrimination.[132] The Court made the grounds for employer liability for harassment clear in a pair

[129] *Id.* (citing Su Lin Blodgett & Brendan O'Connor, *Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English*, PROC. FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACH. LEARNING CONF. (2017), https://arxiv.org/pdf/1707.00061.pdf.
[130] Duarte et al., *supra* note 108, at 15 (citing Rachael Tatman, *Gender and Dialect Bias in YouTube's Automatic Captions*, PROC. OF THE FIRST ASSOC. FOR COMPUTATIONAL LINGUISTICS WORKSHOP ON ETHICS IN NATURAL LANGUAGE PROCESSING 53–59 (2017), http://www. aclweb.org/anthology/W/W17/W17-1606.
[131] Ajunwa, *supra* note 45, at 1685-86.
[132] Meritor Savings Bank v. Vinson, 477 U.S. 57 (1986).

of decisions in 1998, routinely referred to as *Farrager-Ellerth*.[133] At base, the approach to establishing employer liability for sexual harassment depends on the nature of the harassment alleged and the identity of the alleged harasser in relation to the alleged victim.[134]

There are two basic forms of prohibited sexual harassment: (1) quid pro quo harassment and (2) hostile work environment harassment.[135] Quid pro quo harassment typically involves demands that make sexual conduct a condition of employment, meaning that there is an express or implied causal connection between submission to sexually oriented behavior and a tangible job consequence.[136] Hostile work environment harassment, in contrast, involves language or conduct in the workplace that interferes with the employee's ability to work or creates an unbearable work environment.[137] Whether or not an employer is liable for hostile work environment harassment depends to some extent on whether the alleged harasser is a supervisor and hence an agent of the employer or a co-worker of the alleged victim.[138]

In identifying the grounds for employer liability, the Court in *Ellerth* began with the Restatement (Second) of Agency, §219 (2), which dictates that in cases where the agent is operating outside the "scope of employment" as is the case with sexual harassment,[139] the employer is not liable unless:

> (a) the master intended the conduct or the consequences, or
> (b) the master was negligent or reckless, or
> (c) the conduct violated a non-delegable duty of the master, or
> (d) the servant purported to act or to speak on behalf of the principal and there was reliance upon apparent authority, or he was aided in accomplishing the tort by the existence of the agency relation.[140]

In cases involving supervisor harassment, the Court looked to Subsection 219(2)(d) and the "aided in the agency relation" standard to determine that employers may have "vicarious liability" both in cases where there is a tangible employment action taken by the supervisor and in cases without that tangible action subject to the affirmative defense created in the *Farragher/Ellerth* cases.[141]

In contrast, when the alleged harasser is a coworker of the victim, the Court acknowledged that the lower courts have uniformly used the negligence standard to determine employer liability since those cases do not involve a harasser that was aided in the harassment by his or her role as an agent of the employer.[142] As the EEOC explains, with respect to conduct between co-workers, an

---

[133] Faragher v. City of Boca Raton, 524 U.S. 775 (1998); Burlington Industries, Inc. v. Ellerth, 524 U.S. 742 (1998).
[134] *See* Estelle D. Franklin, *Maneuvering Through the Labyrinth: the Employers' Paradox in Responding to Hostile Environment Sexual Harassment - A Proposed Way Out*, 67 Fordham L. Rev. 1517, 15439-48 (1999).
[135] *Id.* at 1540-41
[136] *Id.*
[137] *Id.*
[138] *Id.*
[139] *Ellerth*, 524 U.S. at 756-57 ("The general rule is that sexual harassment by a supervisor is not conduct within the scope of employment.").
[140] *Id.* at 758 (quoting RESTATEMENT (SECOND) OF AGENCY §219(2) (1957)).
[141] *Id.* at 762, 765. ("An employer is subject to vicarious liability to a victimized employee for an actionable hostile environment created by a supervisor with immediate (or successively higher) authority over the employee. When no tangible employment action is taken, a defending employer may raise an affirmative defense to liability or damages.")
[142] *Faragher*, 524 U.S. at 799 (noting that lower federal courts had "uniformly judged employer liability for co-worker harassment under a negligence standard"); Garcez v. Freightliner Corp., 72 P.3d 78, 87 (Or. Ct. App. 2003) (finding that, although the Faragher/Ellerth defense cannot be used in claims of co-worker harassment, its principles are embedded in the requirement that the plaintiff establish that the employer knew or should have known of the harassing conduct (construing Swinton v. Potomac Corp., 270 F.3d 794, 803 (9th Cir. 2001)).

employer is responsible for acts of sexual harassment in the workplace where the employer (or his agents or supervisory employees) "*knew, or should have known* about the harassment and failed to take prompt and appropriate corrective action."[143]

Perhaps the trickiest question in coworker harassment cases centers on the meaning of knowledge—who knew and what does it mean to know? As an initial matter, courts have concluded that there is no need to report to the "correct" supervisor for the negligence standard to apply and that a report to anyone in a supervisory capacity will suffice.[144] In addition, the plaintiff herself need not be the one to put the employer on notice about the co-worker harassment, and the notice need not come through an official report or grievance procedure.[145] "Constructive knowledge" of coworker harassment can support application of the negligence standard without an actual affirmative report by the victim.[146]

However, proving "constructive knowledge" tends to be a steep uphill battle for the plaintiff in a coworker harassment case. In the majority of cases, courts find "constructive notice," without a report by the victim, "only when the acts of harassment are so egregious, numerous, and concentrated as to add up to a campaign of harassment."[147] "Put differently, a victim of coworker harassment must show either actual knowledge on the part of the employer or conduct sufficiently severe and pervasive as to constitute constructive knowledge to the employer.'"[148] Thus, despite allowing for a seemingly employee-friendly "constructive knowledge" standard, courts applying this standard typically demand the existence of blatant and outrageous harassment in order to impute knowledge to the employer.[149] As a result, commentators have referred to this "knew or should have known" standard as an "employer-friendly" one since "it places the burden on the plaintiff to prove that the employer engaged in wrongdoing that renders it liable for the discriminatory harassment."[150] Others have called it the "see-no-evil defence" because "it could mean that the employer is virtually able to ignore the possibility of workplace harassment until it

---

[143] EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, https://www.eeoc.gov/harassment (last visited May 20, 2021).
[144] Varner v. Nat'l Super Mkts., Inc., 94 F.3d 1209, 1213 (8th Cir. 1996) ("Indeed, Title VII defines employer to include any 'agent' of the employer." (citing 42 U.S.C. § 2000e(b)).
[145] *See, e.g.,* Flowers v. Honigman Miller Schwarz & Cohn, No. 04-71928, 2005 U.S. Dist. LEXIS 9062, at *22 (E.D. Mich. May 16, 2005) (noting that "it is not necessary that the plaintiff herself put the employer on notice" when the employer knew or should have known of the harassment) (cited in Anne Lawton, *The Bad Apple Theory In Sexual Harassment Law*, 13 GEO. MASON L. REV. 817, 847, n.162 (2006)).
[146] *See, e.g.*, Blakey v. Cont'l Airlines, 164 N.J. 38, 45 (2000) (considering whether "an employer, having actual or constructive knowledge that co-employees are posting harassing, retaliatory, and sometimes defamatory, messages about a co-employee on a bulletin board used by the company's employees, [has] a duty to prevent the continuation of such harassing conduct[.]").
[147] Rennard v. Woodworker's Supply, Inc., 101 F. App'x 296, 304 (10th Cir. 2004). *See also* Watson v. Blue Circle, Inc., 324 F.3d 1252, 1259 (11th Cir. 2003) (citation omitted) (holding that "constructive notice is established when the harassment was so severe and pervasive that management reasonably should have known of it"); Sharp v. City of Houston, 164 F.3d 923, 930 (11th Cir. 1999) (explaining that an employer has constructive knowledge of co-worker harassment only when someone "with remedial power over the harasser . . . knew or should have known" of the harassing conduct) (cited in Lawton, *supra* note 145, at 847 n.162).
[148] Miranda Oshige McGowan, *Certain Illusions About Speech: Why the Free-Speech Critique of Hostile Work Environment Harassment Is Wrong*, 19 CONST. COMMENT. 391, 439-40 (2002) (quoting Miller v. Kenworth of Dothan, Inc., 277 F. 3d 1269, 1278 (11th Cir. 2002)).
[149] *Id.* "[A]bsent notification, courts will find constructive knowledge only if harassment is public and constant, or if the acts of harassment are 'so egregious, numerous, and concentrated' that they amount to a 'campaign'") (citations omitted).
[150] Lu-in Wang, *When The Customer Is King: Employment Discrimination As Customer Service when The Customer Is King: Employment Discrimination As Customer Service*, 23 VA. J. SOC. POL'Y & L. 249, 260 (2016).

is reported."[151] As a result, it is nearly impossible to satisfy the negligence standard of liability absent a formal or informal report to management.[152]

How then does this employer-friendly standard fare when applied to a case in which the employer was actively monitoring employee communications for possible instances of inappropriate or harassing conduct? In short, I propose – not well. Commentators have begun to theorize about the impact of social media on the negligence standard for coworker harassment. Imagining a scenario in which an employee posts harassing comments on Facebook about another employee and a manager who is "friends" with the harasser has access to this content, the argument is as follows:

> If the manager happens to see harassing comments or offensive dialogue being exchanged among employees and does nothing, the employer may later be accused of having knowledge of co-worker harassment and failing to respond. Even if the manager never actually saw the inappropriate content, the existence of the friendship with the subordinate in and of itself may give an enterprising plaintiff-employee the opportunity to claim that the employer knew *or should have known* of the alleged harassment.[153]

When employers utilize #MeTooBots to monitor employee communications, they essentially forego the "knew or should have known" defense that would have prevented liability absent an employee report. Instead, by expressly claiming to monitor all employee communications to detect inappropriate behaviors and harassment, they are adopting a system that creates the very knowledge that the negligence standard requires.[154] At the very least, a plaintiff should be able to convincingly argue that the employer "should have known" of the harassment since it employs a tool that promises to unearth it.

---

[151] Angela Duffy, *Comparative Analysis of the Vicarious Liability of Employers in Harassment Cases in the United Kingdom and the United States*, COMMON L. WORLD REV. (CLWR) 31 3 (254) (2002).

[152] *See* Rennard v. Woodworker's Supply, Inc., 101 F. Appx. 296, 304 (stating that the employer has actual notice of harassment when the victim reports to "management-level employees").

[153] Laura Thalacker & Kelly Kichline, *Special Feature: Pitfall Potential: The Risks of Social Media*, 18 NEVADA LAWYER 16, 17-19 (2010). This type of argument has begun to appear in other contexts including race discrimination claims under Title VI and tort cases involving online child pornography. *See e.g.,* Stafford v. George Wash. Univ., No. 18-cv-2789 (CRC), 2019 U.S. Dist. LEXIS 94088, at *4 (D.D.C. June 5, 2019) (considering constructive knowledge of environment of harassment and discrimination, in part, because "the head coach, Munoz, "was aware of these racist postings because he was Facebook friends with the tennis players and told the players that he monitored social media postings[.]"); Doe v. XYC Corp., 887 A.2d 1156, 1159 (N.J. Super. Ct. App. Div. 2005) (imputing knowledge to defendant employer that employee was using his work computer to access child pornography since an investigation of his computer (that should have been triggered by reports) would have revealed the employee's activities).

[154] In the one recent case that dealt with co-worker harassment in an online bulletin board, the court, although focused primarily on whether the bulletin board was a part of the workplace, had this to say about the potential for negligence liability for co-worker posts: "To repeat, employers do not have a duty to monitor private communications of their employees; employers do have a duty to take effective measures to stop co-employee harassment when the employer knows or has reason to know that such harassment is part of a pattern of harassment that is taking place in the workplace and in settings that are related to the workplace." *Blakey*, 164 N.J. at 62, 751 A.2d at 552 (2000). *See also* Donald P. Harris, Daniel B. Garrie, Matthew J. Armstrong, *Sexual Harassment: Limiting the Affirmative Defense in the Digital Workplace*, 39 U. MICH. J.L. REFORM 73 (2005) (arguing that in cases where employers have the capacity to monitor employee communications, the affirmative defense should be modified to allow courts to "explore whether or not the defendant took reasonable steps to monitor and block the offensive digital communications or whether the defendant can articulate a rational reason for not utilizing readily available technology to prevent workplace digital harassment.").

However, as described above, the AI-based tools on which employers rely for this monitoring are often functionally incapable of the task, may be plagued by signal and bias problems, and, without a doubt fail to achieve a 100 percent accuracy rate. While the danger that the AI-based monitor will flag benign communications causing problems for all of the employees involved is a real one, there is also the significant danger that the tool fails to flag significant harassment that it should have captured.

Imagine a case in which an employee routinely receives sexual emails from a co-worker. The emails contain images and words but using vernacular or a language on which the AI is not trained. The recipient of the emails is distressed and embarrassed. Her work begins to suffer as she grows fearful of opening her email and of seeing the sender in person. She, however, does not report these communications to anyone both because she is uncomfortable discussing them openly and because she assumes that management is aware of them and unconcerned because she knows about the AI-based monitoring and has received no word of a company investigation or the like. In that case, the plaintiff could reasonably argue that an affirmative report of the harassment to a manager is unnecessary to impute knowledge and ultimately liability. She suffered under a coworker-created hostile work environment that the employer should have known about and yet nothing was done to stop it. Absent the #MeTooBot, the employer could have easily argued that it had no knowledge (actual or constructive) of the harassment and thus could not be liable. Use of the AI-based monitor turns a case that would have been an almost automatic dismissal under the negligence standard into a very problematic legal case for the employer.

2. Impact on Employee Protection Against Retaliation

Much like the potential for increased liability for employers using AI-based harassment monitors, the unintended legal consequences of #MeTooBots may be equally problematic for employees who are victims of harassing communications (despite being the technology's raison d'être). Title VII and its related doctrine generally prohibit retaliation against victims of unlawful harassment who report the conduct to management along with bystanders who take an active role in opposing such unlawful conduct. However, absent some form of "opposition" or "participation" conduct, a victim of harassment receives no legal protection against retaliation.[155] When the human reporter is removed from the process of identifying harassment, the victim may lose both her voice and agency in the process as well as her protection against retaliatory conduct.

Imagine the following scenario: A company's AI-based monitor identifies inappropriate language (per the rules on which it was trained) in several email communications between a supervisor and an employee. The bot reports the communications to Human Resources, which commences an investigation. Management ultimately disciplines the supervisor by sending him to sexual harassment training and moving him to a different department. His team is extremely upset, having seen him as a great manager and mentor. During the investigation process, the victim's name is leaked. Without knowing much else about the incident, the victim's team turns on her. She is shunned by co-workers, left off team e-mails and meetings, and generally ignored. She gains a reputation as a "complainer," which, in turn, may ultimately impact her future assignments and even promotions. Such behavior might seem to be obviously in violation of Title VII's prohibition on retaliation but, as this Section will describe, because of the way retaliation protection is structured, the victim in this scenario may be treated by courts as legally helpless.

[155] *See infra* text accompanying notes 160-70.

Title VII prohibits retaliation as follows:

> It shall be an unlawful employment practice for an employer to discriminate against any of his employees or applicants for employment . . . because he has opposed any practice made an unlawful employment practice by this subchapter, or because he has made a charge, testified, assisted, or participated in any manner in an investigation, proceeding, or hearing under this subchapter.[156]

The statute makes retaliation protection available for two types of conduct: "opposition conduct" and "participation conduct."[157] "Opposition conduct" includes internal complaints and can include complaints that are written or verbal, formal or informal, proactively made by an employee or in response to a question by management.[158] In contrast, "participation conduct" pertains exclusively to an employee's participation in an investigation by the EEOC, a proceeding in court, or the employee's own filing of charges or suit.[159] To prevail on a retaliation claim, the plaintiff must demonstrate that he (1) engaged in a statutorily protected activity or expression (e.g. a complaint of harassment); (2) that he suffered an adverse action by the employer (e.g. promotion denial); and (3) that there was a causal link between the protected action or expression and the adverse action (the short time between complaint and termination often provides this causal link without any other evidence of causation).[160] In the scenario laid out above (and in most #MeTooBot cases), the "Participation clause" is irrelevant because there has been no charge or lawsuit filed and the sole investigation was internal. The victim would thus have to rely on the "opposition clause" and attempt to demonstrate some form of opposition activity. Herein lies the problem in AI-discovered cases.

The Supreme Court in recent years has dramatically expanded the meaning of "opposition conduct" under Title VII and other related employment statutes. In 2008, the Court endorsed the conclusion that 42 U.S.C. §1981 includes an implied prohibition on retaliation in a case involving a plaintiff who faced retaliation because he complained to managers about the dismissal of a coworker,[161] impliedly concluding that "opposition conduct" includes complaints about another

---

[156] 42 U.S. Code § 2000e–3(a).

[157] *See* Leora Eisenstadt and Deanna Geddes, *Suppressed Anger, Retaliation Doctrine, and Workplace Culture*, 20 UNIVERSITY OF PENNSYLVANIA JOURNAL OF BUSINESS LAW 147, 156 (2018) (citing Crawford v. Metro. Gov't of Nashville & Davidson County, 129 S. Ct. 846, at 850 (2009) ("[t]he Title VII anti-retaliation provision has two clauses, making it an unlawful employment practice for an employer to discriminate against any of his employees . . . [1] because he has opposed any practice made an unlawful employment practice by this subchapter, or [2] because he has made a charge, testified, assisted, or participated in any manner in an investigation, proceeding, or hearing under this subchapter.' The one is known as the 'opposition clause,' the other as the 'participation clause.'")).

[158] *See Crawford*, 129 S. Ct. at 850. *See also* Kasten v. Saint-Gobain Performance Plastics Corp., 563 U.S. 1, 15 (2011) (finding that the Fair Labor Standards Act applies to a complaint, whether oral or written).

[159] See *Crawford*, 129 S. Ct. at 850 (describing the process by which an employee may bring a statutory claim against an employer).

[160] Eisenstadt and Geddes, *supra* note 157, at 157 (citing Lawrence D. Rosenthal, *Timing Isn't Everything: Establishing a Title VII Retaliation Prima Facie Case After University of Texas Southwestern Medical Center v. Nassar*, 69 SMU L. REV. 143, 152 (2016); EEOC v. Avery Dennison Corp., 104 F.3d 858, 860 (6th Cir. 1997) (holding that "to find a prima facie case of retaliation under Title VII, a plaintiff must prove by a preponderance of the evidence: 1) plaintiff engaged in activity protected by Title VII; 2) plaintiff's exercise of his civil rights was known by the defendant; 3) that, thereafter, the defendant took an employment action adverse to the plaintiff; and 4) that there was a causal connection between the protected activity and the adverse employment action.").

[161] CBOCS W., Inc. v. Humphries, 553 U.S. 442, 445, 128 S. Ct. 1951, 1954 (2008).

employee's mistreatment, not only your own.[162] One year later, the Court's decision in *Crawford v. Metropolitan Government of Nashville & Davidson County* concluded that "opposition conduct" also includes responding to questions as part of an internal investigation as opposed to only encompassing affirmative complaints.[163] The Court pragmatically noted that "[n]othing in the statute requires a freakish rule protecting an employee who reports discrimination on her own initiative but not one who reports the same discrimination in the same words when her boss asks a question."[164] Finally, in 2011, in a Fair Labor Standards Act (FLSA) case, the Court concluded that affirmative complaints were protected activity for retaliation purposes whether they were in written or verbal form.[165]

What the Court has not done, however, is expand "opposition conduct" to uncommunicated or silent opposition. The only real discussion of the subject by the Court can be found in Justice Alito's concurrence in *Crawford* in which he wrote specifically to oppose the extension of the doctrine beyond the facts of the case before the Court: "the Court's holding does not and should not extend beyond employees who testify in internal investigations or engage in analogous purposive conduct."[166] In laying out his position, Justice Alito examined Random House's definitions of the term "oppose," noting that most of the definitions include the taking of affirmative action, but that the fourth definition, which the Court mentioned in dicta, "goes further, defining 'oppose' to mean 'to be hostile or adverse to, as in opinion.' Thus, this definition embraces silent opposition."[167] Justice Alito wrote to clarify that the Court's decision should not extend to silent opposition and cautioned that such an expansion would have problematic implications.[168]

A number of commentators have explored this question in detail. In commenting on the *Crawford* decision, Alex Long concludes that the Court is generally looking for affirmative action directed to the employer when it considers "opposition conduct."

> Nearly all of the examples the Court used in attempting to explain what qualifies as "opposing" unlawful conduct involved expressions of disapproval communicated directly to the employer. From providing a disapproving account to the employer of an employee's behavior, to communicating a belief to the employer that the employer's conduct was unlawful, to refusing to obey the employer's orders, the examples the Court relied upon involved expressions of disapproval. . . . [t]he Court's sympathy is premised on the assumption that the employee expresses her belief to the employer that discrimination has occurred. This is certainly the view of Justices Alito and Thomas, who concurred in the judgment but argued that the term "opposition" should include only "active and purposive" conduct. In sum,

[162] The lower courts have followed suit. In 2010, the First Circuit concluded that opposition conduct includes accompanying another employee to Human Resources to make a complaint and speaking to the alleged harasser on someone else's behalf. Collzao v. Bristol-Myers Squib Mfg., 617 F.3d 39, 47 (1st Cir. 2010).
[163] 129 S. Ct. 846 (2009).
[164] *Id.* at 851.
[165] *Kasten*, 563 U.S. at 14 ("To fall within the scope of the antiretaliation provision, a complaint must be sufficiently clear and detailed for a reasonable employer to understand it, in light of both content and context, as an assertion of rights protected by the statute and a call for their protection. This standard can be met, however, by oral complaints, as well as by written ones.").
[166] *Crawford*, 555 U.S. at 281 (Alito, J. concurring).
[167] *Id.* at 282.
[168] *Id*. at 282-83.

> Crawford does not provide much comfort to employees who do not openly express disapproval of an employer's actions to the employer.[169]

Matt Green agrees with this assessment of the Court's view in *Crawford*, and notes that this reading of the provision "as requiring some form of expression versus merely holding an opinion is consistent with the Supreme Court's characterization of a retaliation claim under Title VII as a conduct-based claim."[170] "[U]nlike the substantive provisions of the statute, which protect individuals because of who they are, i.e., their status, the anti-retaliation provision 'seeks to prevent harm to individuals based on what they do, i.e., their conduct.'"[171]

Nevertheless, Green argues for a broader standard for "opposition conduct" than Alito and Thomas's "purposive conduct." He proposes a hypothetical in which a victim of harassment who does not want to report it to management instead vents to a coworker. The coworker then reports the conduct to management. If the victim is later terminated in retaliation for venting to a coworker, that adverse action would nevertheless be lawful under the "purposive conduct" standard.[172] Insisting that the venting to the coworker should be protected conduct, Green argues that "[w]hat matters is whether the opposition reaches the employer's attention and whether the employer discriminates on the basis of the opposition."[173] He would thus expand the meaning of "opposition conduct" to include affirmative action that is indirectly or unintentionally communicated to the employer.

Unfortunately for victims of AI-detected harassment, neither the Alito standard nor Green's proposed expanded standard would likely offer protection. Under Alito's "purposive conduct" approach, the victim of AI-detected harassment in our hypothetical is devoid of options—she may or may not have been the victim of actionable harassment but importantly, she has done nothing to express her opposition to the underlying conduct. In fact, she may not even have had the chance to express opposition before the AI monitor discovered and reported the inappropriate communications. That she faced ridicule for her time with potentially long-term impacts on her career may not help her case. If the court is looking for conduct-based opposition, this victim will find herself unprotected by the retaliation doctrines.[174]

Somewhat less obvious is the victim's fate under Green's proposed expanded standard, which would include opposition that is "indirectly and unintentionally expressed to an employer."[175] While this is a common sense expansion, it would still likely leave the victim of AI-detected harassment unprotected.[176] In our scenario, the victim has not engaged in any conduct at all. Her

---

[169] Alex Long, *Employment Retaliation and the Accident of Text*, 90 OR. L. REV. 525, 556-557 (2011).
[170] Matthew Green, *Express Yourself: Striking a Balance Between Silence and Active, Purposive Opposition Under Title VII's Anti-Retaliation Provision*, 28 HOFSTRA LAB. & EMP. L.J. 107, 126 (2010).
[171] *Id.* at 127.
[172] *Id.* at 110-11.
[173] *Id.* at 111.
[174] As Green points out, the lower courts have largely endorsed this "purposive conduct" standard as well. *See* Green, *supra* note 170, at 133-34 (discussing Pitrolo v. County of Buncombe, No. 07-2145, 2009 WL 1010634 (4th Cir. Mar. 11, 2009) and Ackel v. National Communications, 339 F.3d 376 (5th Cir. 2003) (rejecting retaliation protection to plaintiffs who complained to family members or non-supervisory employees who then reported to management)).
[175] Green, *supra* note 170, at 111.
[176] There is a second, ironic way in which the AI makes victims of harassment more vulnerable to retaliation. Under the doctrine known as the "Objectively Reasonable Belief" doctrine, endorsed in the U.S. Supreme Court's 2001 decision in Clark County School District v. Breeden, 532 U.S. 268, 270–71 (2001) (per curiam), for retaliatory conduct to be unlawful, the complaining party must have an objectively reasonable belief that the practices he or she opposed (which gave rise to the retaliation) were unlawful. *See* Boles et al. *supra* note 115, at 168-69. When

opposition (if she opposes the conduct at all) is silent at the point in time when her name is leaked. She may not have wanted to report the conduct to management, to a coworker or even to a family member. Nevertheless, she was the victim of the technology's ability to detect and report without her cooperation at all.

It is possible that a thoughtful, nuanced court may view the totality of circumstances in these cases and find a means of protecting the AI-detected harassment victim. For example, although the prohibition on retaliation is discussed in a separate provision of Title VII, the underlying harassing comments detected by the AI along with the victim's treatment after her name was leaked may all be considered to be part of the unlawful harassment.[177] As the EEOC notes, "Harassment becomes unlawful where 1) *enduring the offensive conduct becomes a condition of continued employment*, or 2) the conduct is severe or pervasive enough to create a work environment that a reasonable person would consider intimidating, hostile, or abusive."[178] Moreover, the EEOC identifies that harassment includes, "offensive jokes, slurs, epithets or name calling, physical assaults or threats, intimidation, ridicule or mockery, insults or put-downs, offensive objects or pictures, and *interference with work performance*" and the victim of harassment may be "anyone affected by the offensive conduct" even if there is no "economic injury to, or discharge of, the victim."[179] In the hypothetical above, the victim of inappropriate digital communications, even if not themselves unlawful, may be understood to be a victim of unlawful harassment once she is forced to endure the ostracism and negative treatment associated with her name being leaked as part of the investigation. A skilled plaintiff's lawyer may be able to frame the subsequent retaliatory conduct as harassment in and of itself without relying on the retaliation provision and its narrow approach that protects only those who have taken purposeful action to protest the harassment.

Unfortunately, the thoughtful and nuanced court needed to make this leap is increasingly rare. As a number of scholars have highlighted, many courts, focused on the numerous doctrines created to assess claims of harassment and retaliation become mired in parsing the technical details of the claims and essentially miss the forest for the trees.[180] As a practical matter, it is far

---

deciding whether that belief was "objectively reasonable," the court does not consider the "good faith" or actual belief of the reporter but rather to whether a court would consider the reported behavior to be unlawful discrimination or harassment. *Id.* (citing *Breeden,* 532 U.S. at 270-71; Satterwhite v. City of Houston, 602 F. App'x 585, 588 (5th Cir. 2015)). When applied to a case of AI-detected harassment, an obvious problem arises. If the AI-based monitor flags and reports communications that it is trained to view as inappropriate but that do not rise to the level of unlawful harassment (which may happen because the communications are not sufficiently "severe or pervasive" to meet the standard), and the victim of the inappropriate communications is nevertheless retaliated against as a result of the AI's flag and resulting investigation, the victim cannot claim unlawful retaliation. If the underlying conduct does not meet the standard for unlawful discrimination as interpreted by a court, the resulting retaliation is perfectly lawful.

[177] *See* Brake, supra note 12, at 21 (contending that "[r]ecognizing retaliation as a form of discrimination, one that is implicitly banned by general proscriptions of discrimination, pushes the boundaries of dominant understandings of discrimination in useful and productive ways."); *see also* Brief of Employment Law Professors as Amici Curiae in Support of Respondent at 5, Univ. of Tex. Southwestern Med. Ctr. v. Nassar, 133 S. Ct. 2517 (2013) (No. 12-484) (contending "[a] long line of cases confirms that when Congress uses the word "discriminate" that term encompasses retaliation.").

[178] EEOC, *supra* note 70 (emphasis added).

[179] *Id.*

[180] *See* Sandra Sperino, *Into the Weeds: Modern Discrimination Law*, 95 NOTRE DAME L. REV. 1077, 1078-79 (2020) (critiquing the overreliance on court-created frameworks, roadmaps, and "ancillary doctrines or subdoctrines" that capture "an ever-increasing amount of judicial attention" as being ineffective and "underestimat[ing] the complexity of the modern workplace."). *See also* Marcia McCormick, *Let's Pretend that*

more likely that both the plaintiff's attorney and the court reviewing the claims will identify the negative treatment as retaliation, becoming entangled in the doctrines on opposition and participation conduct and finding that the silent victim of AI-detected harassment falls between the cracks of these doctrines.

3. Impact on Victim Voice

Beyond exposing victims of harassment to "lawful" retaliation, AI-based harassment monitors create significant ethical problems with respect to the victim's agency and voice. In their recent article on "hushing contracts," David Hoffman and Erik Lampmann explore the use of non-disclosure agreements in sexual misconduct cases and describe one of the costs of such agreements as "the deprivation of survivors' ability to openly and honestly talk about their experiences" or, in essence, their ability to "'come out' as survivors."[181] The use of #MeTooBots, however well-intended, creates the equally problematic flip side of this harm—it deprives victims of the ability *not to speak*.

In the typical workplace sexual harassment case, a man or woman may be the victim of multiple instances of harassing comments or conduct over a period of time or one particularly horrible instance before he or she decides to report the behavior. Either way, it is likely an agonizing decision. In addition to the fear of concrete retaliation in the form of demotion, denial of opportunities, or outright termination, the victim fears losing the reputation she has worked hard to build, being viewed as a complainer or problem employee, forfeiting political capital she had intended to use elsewhere, and being seen henceforth through the lens of this complaint rather than the workplace successes she has achieved.[182] The choice to speak up may be based on her inability to function in the workplace otherwise, her desire to help others avoid such treatment, or her desire to reclaim the power that the harassment has sought to steal from her.[183]

Likewise, the choice to remain silent may be a deliberate one, based on knowledge born of watching those who spoke out before. It may result from life circumstances that make it impossible

---

Federal Courts Aren't Hostile to Discrimination Claims, 76 OHIO ST. L.J. 22, 28-29 (2015) (describing "decisionmaking heuristics for employment discrimination cases" that lead to increasingly narrow rulings); Kerri Lynn Stone, *Shortcuts in Employment Discrimination Law*, 56 ST. LOUIS L.J. 111, 113-114 (2011) (discussing the way in which courts rely on "shortcut" doctrines that unfairly skew towards the defendant "at the expense of a more holistic assessment of all properly-considered evidence against the backdrop of the overarching question posed by the relevant legislation: did employment discrimination "because" of an unlawful consideration occur?").

[181] David Hoffman & Erik Lampmann, *Hushing Contracts*, 97 WASH. U. L. REV. 165, 179 (2019).

[182] *See* Joanna L. Grossman, *The Culture of Compliance: The Final Triumph of Form Over Substance in Sexual Harassment Law,* 26 HARV. WOMEN'S L.J. 3, 75 (2003) (describing numerous reasons why victims fail to report harassment, including fear of retaliation, ostracization, and alienation from mentors and coworkers, "or because 'calling attention to offensive behavior reinforces stereotypes of women as victims.'" (quoting Nina Burleigh & Stephanie B. Goldberg, *Breaking the Silence: Sexual Harassment in Law Firms*, A.B.A. J. 46, 48 (Aug. 1989) and citing Denise H. Lach & Patricia A. Gwartney-Gibbs, Sociological Perspectives on Sexual Harassment and Workplace Dispute Resolution, 42 J. VOCATIONAL BEHAV. 102, 111 (1993) (describing survey data about job consequences for filing sexual harassment complaints); Jan Salisbury et al., Counseling Victims of Sexual Harassment, 23 PSYCHOTHERAPY 316, 319 (1986) (noting that the "occurrence of physical and mental symptoms is dramatically higher [for those who file formal complaints] then [sic] for those who do not").

[183] *See* Vicki Schultz, *Reconceptualizing Sexual Harassment, Again*, 128 YALE L.J.F. 22, 24 (2018); Vicki Schultz, *Open Statement on Sexual Harassment from Employment Discrimination Law Scholars*, 71 STAN. L. REV. ONLINE 17 (2018), https://www.stanfordlawreview.or g/online/open-statement-on-sexual-harassment-from-employment-discrimination-law-scholars/ [https:/ /perma.cc/DH55-D6TQ].

to risk job loss or income reduction.[184] It may be a considered response to the beginnings of a campaign of harassment that are not yet severe enough to be taken seriously by the employer or courts but if left to fester, might be actionable and better to report later.[185] It might even be a recognition of the power the victim has in the organization and the desire to maintain that power to assist other women and disadvantaged groups.[186] Regardless of the *why*, the act of staying silent is, without doubt, an act of agency or self-direction. The use of #MeTooBots to monitor, detect, and report harassment deprives victims of this agency. Ironically, again, an effort to help (typically female) victims of harassment, this technology visits upon victims another abuse. As Linda Mills notes, "Feminist political practice - even in the name of gender warfare - should not mimic patriarchy through either the use of threat tactics or the inattention to individual desire."[187]

To explain the abuse of voice deprivation, I turn to the literature on domestic violence and sexual assault and the rise and critique of mandatory referral and mandatory arrest and prosecution laws to combat the failure to take such offenses seriously. In recent years, spurred by campus activism around gender violence at colleges and universities, there have been a "series of 'mandatory referral' laws proposed in state legislatures and the United States House of Representatives."[188] Responding to the critique that schools fail to properly investigate and adjudicate claims of sexual violence, these laws mandate referral of school reports of such violence to local law enforcement for prosecution.[189] This recent effort parallels a similar approach in domestic violence law. In the late 1970's and 1980's, in response to case after case of police ignoring horrific abuse of women by their husbands, in jurisdictions "throughout the country, either legislatures imposed or police departments implemented policies requiring arrests in domestic violence cases whenever police had probable cause to do so."[190] In addition, jurisdictions began to respond to the problem of prosecutorial discretion, since "prosecutors had also routinely chosen not to pursue cases against the few perpetrators of violence who police had actually

---

[184] *See* Ben Bursten, *Psychiatric Injury in Women's Workplaces*, 14 BULL. AM. ACAD. PSYCHIATRY L. 245, 248 (1986) ("[The] social fact that women need employment that may not be abundantly available tends to create a willingness to tolerate persistently abusive conditions of work.").

[185] *See* Lawrence D. Rosenthal, *To Report or Not to Report: The Case for Eliminating the Objectively Reasonable Requirement for Opposition Activities Under Title VII's Anti-Retaliation Provision*, 39 ARIZ. ST. L.J. 1127, 1158 (2007) ("One of the biggest problems with maintaining the objectively reasonable standard in these cases is that it puts employees in the unenviable position of having to decide whether to report an offending co-worker and run the risk of termination, or keep quiet and run the risk of having to endure working in a hostile environment, regardless of whether that environment meets the Harris standard for actionable harassment.").

[186] *See* Heather McLaughlin, Christopher Uggen, and Amy Blackstone, *Sexual Harassment, Workplace Authority, and the Paradox of Power*, 77(4) AM. SOC. REV. 625, 641 (2012) (describing findings that "female supervisors are more, rather than less, likely to be harassed . . . Although women supervisors' authority is legitimated by their employer, sexual harassment functions, in part, as a tool to enforce gender-appropriate behavior. . . . When women's power is viewed as illegitimate or easily undermined, co-workers, clients, and supervisors appear to employ harassment as an 'equalizer' against women supervisors, consistent with research showing that harassment is less about sexual desire than about control and domination.").

[187] Linda Mills, *Killing Her Softly: Intimate Abuse and the Violence of State Intervention*, 113 HARV. L. REV. 550, 568 (1999).

[188] Alexandra Brodsky, *Against Taking Rape "Seriously": The Case Against Mandatory Referral Laws for Campus Gender Violence*, 53 HARV. C.R.-C.L. L. REV. 131, 133, 139 n.46-49 (2018) (citing Safe Campus Act, H.R. 3404, 114th Cong. (2015) and describing bills in at least eleven states, four of which have become law).

[189] *Id.* at 138-40.

[190] Leigh Goodmark, *Autonomy Feminism: An Anti-Essentialist Critique of Mandatory Interventions in Domestic Violence Cases*, 37 Fla. St. U.L. Rev. 1, 3 (2009). *See also* Aya Gruber, *Rape, Feminism, and the War on Crime*, 84 WASH. L. REV. 581, 649-650 (discussing feminist scholars critique of mandatory prosecution and similar policies that "often inure to the great detriment of abuse survivors").

arrested."[191] In response, they implemented "no-drop prosecution" policies in which "prosecutors would not dismiss criminal charges in otherwise winnable cases simply because the victim was not interested in, or was even adamantly opposed to, pursuing the case."[192]

Much like the development of #MeTooBots, the enactment and implementation of these laws and policies aim to protect victims (most often women) who suffer abuse of some kind and who the system has failed to protect. "Mandatory referral" laws endeavor to help institutions identify and punish wrong-doers, identify victims to provide resources and support, and collect data that fuels future policies aimed at deterrence.[193] "Mandatory arrest" laws similarly seek to keep women safe.[194] The approach in all these cases to the problem of a failed system, however, is to supplant the victim's agency and decision-making power with that of the system itself. And, as is also the case in workplace sexual harassment, the result of these well-meaning interventions is to visit a different kind of abuse on those they seek to protect.

The critiques of "mandatory referral," "mandatory arrest," and "no-drop prosecution" policies center on the impact these approaches have on victim voice and agency.[195] In the case of "mandatory referral" rules for campus sexual assault, critics argue that the unintended consequences can be both physical and psychological. "For students who disclose their victimization, these policies infringe their autonomy (that is, self-determination) and aggravate the psychological and/or physical harm caused by the violence itself."[196] The ability to control one's fate is essential to identity, to a victim's self-confidence, and ability to perform elsewhere in her life. "An institutional response that overrides the survivor's own preferences also calls into question a woman's judgment, and thereby produces additional harm."[197] In addition, when the system deprives victims of agency, it assumes that the same response is equally effective in all cases regardless of the context, the victim and perpetrator's identities and roles, and the like. "Creating space for choice honors the differences between women, recognizing that race, class, sexual orientation, disability status, and a multiplicity of other variables color how a particular woman might want to respond to a particular incidence of violence at a particular moment in time."[198] In the context of domestic violence cases, critics of "mandatory arrest and prosecution" rules highlight the impact of the loss of autonomy on the victim's healing. "No intervention that takes power away from the survivor can possibly foster her recovery, no matter how much it appears to be in her immediate best interest."[199] This revictimization can be psychological and

[191] *Id.* at 10.
[192] *Id.* at 11-12.
[193] Merle H. Weiner, *A Principled and Legal Approach to Title IX Reporting*, 85 Tenn. L. Rev. 71, 82-83 (2017).
[194] Mills, *supra* note 187, at 563-64 ("[A]t one level, the system clearly colluded with the batterer and replicated the violence endemic to patriarchy by failing to take the victim's complaints seriously. . . .Many advocates of mandatory arrest and prosecution have argued that mandatory policies force state actors to treat crimes against women in the same manner in which they treat other crimes.").
[195] While some critics use the terms autonomy and agency interchangeably, Goodmark cautions that the term "autonomy" itself carries some "philosophical baggage" because women "by virtue of their subordinated status as victims of a patriarchal system, are rarely able to exercise the sort of autonomy contemplated by philosophers" and notes that some have instead chosen to use "agency, which captures the key features of autonomy-self-definition and self-direction-but recognizes how social construction delimits the choices available to women." Goodmark, *supra* note 190, at 24 (quoting Kathryn Abrams, *From Autonomy to Agency: Feminist Perspectives on Self-Direction*, 40 WM. & MARY L. REV. 805, 823-24 (1999)).
[196] Weiner, *supra* note 193 at 87.
[197] *Id.* at 93.
[198] *Id.* at 90.
[199] Mills, *supra* note 187, at 577.

physical. "[B]ypassing a victim's assessment of her own needs, safety and otherwise, can be lethal."[200]

Although there are numerous differences between the criminal justice system's approach to domestic violence and employers' responses to harassment, many of these critiques apply to the use of AI-based monitors for workplace sexual harassment. As one scholar notes in the domestic violence context, "If, as most scholars agree, domestic violence is characterized by a power imbalance between the parties, restoring power to women who have been battered should be a priority when crafting domestic violence law and policy."[201] Similarly, sexual harassment is itself an expression of power.[202] An employer's investigation of possible harassment should acknowledge this and not impose a secondary deprivation of power on its victims. The loss of agency in this context, as with campus sexual assault and domestic violence, ignores the differences between individual victims and between instances of sexual harassment. It has the potential to cause even greater harm both psychologically and financially. A woman who chooses not to report harassment in the workplace may be considering a multitude of factors in making that choice, and the AI-based detector should not supplant her judgment.[203]

In Ifeoma Ajunwa's recent article exploring AI's potentials and pitfalls in combating bias, she urges consideration of the sometimes unpredictable ways in which the technology impacts society.

> What elements of the social world does a new technology make particularly salient that went relatively unnoticed before? What features of human activity or of the human condition does a technological change foreground, emphasize, or problematize? And what are the consequences for human freedom of making this aspect more important, more pervasive, or more central than it was before?[204]

Here, the technology does not emphasize an unnoticed element; instead it eliminates one—the human reporter. Aiming to respond to the problem of fear-based underreporting of sexual harassment, the technology replaces her role entirely. Unfortunately and ironically, in doing so, it leaves the victim of harassment potentially unprotected against retaliatory conduct and deprives her of the choice to speak up or stay silent, visiting an additional abuse on already victimized individual. In attempting to alleviate the harm caused by workplace harassment, the technology inadvertently exposes victims to more of it.

---

[200] Rebecca Fialk and Tamara Mitchel, *Jurisprudence: Due Process Concerns For The Underrepresented Domestic Violence Victim*, 13 BUFF. WOMEN'S L.J. 171, 206 (2005).

[201] Goodmark, *supra* note 190, at 29.

[202] *See generally* Schultz, *Reconceptualizing Sexual Harassment, Again*, *supra* note 183.

[203] Notwithstanding this problem of victim voice in the harassment context, I acknowledge that the differences in context between the criminal justice system and the workplace are sufficiently large that a different calculus may be necessary. For example, although police and prosecution policies and practices may enable domestic violence and sexual assault to some extent, those institutions do not exert the same level of day-to-day control over victims' lives as employers exert over their employees. In contrast, the employment relationship is what provides the opportunity and context for an employee to be subject to sexual harassment and, as such, the employer may bear a greater burden in its eradication. In addition, the employer needs to consider not only the employee who is the target of the particular instance of harassment but also the numerous other employees who may be negatively impacted by it either because they also operate within the hostile work environment created by the harasser or because the victim, while choosing not to report to management, may share her anger, fear, and perception of injustice with her co-workers, creating a culture of discontent. *See* Eisenstadt & Geddes, *supra* note 157, at 182-86. As a result, it is plausible that an employer may be more concerned with the overall workplace culture implications of a failure to identify instances of harassment than it is with a single employee's agency and power in choosing to report or stay silent.

[204] Ajunwa, *supra* note 45, at 1675.

4. Impact on Organizational Culture

As is often the case with AI in the workplace, the intention behind its use and the results in reality are not always aligned. Ajunwa refers to this as the "Paradox of Automation."[205] In the hiring context, "the trend of hiring by algorithm grew out of a cottage industry of tech start-ups seeking to help diversify Silicon Valley" by aiming to reduce the power of implicit bias, which is often seen as an unavoidable component of human decision-making.[206] However, as Ifeoma Ajunwa, Stephanie Bornstein, and others point out, the use of AI in hiring and workplace management, in fact, ends up undermining the anti-discrimination protections that are available.[207] As is described in the prior sections, the same is true for #MeTooBots—the technology is marketed as a means of protecting against the problems of workplace sexual harassment but can end up exposing employers to a heightened risk of liability; it is promoted as a solution to the underreporting of harassment due to fear of retaliation but can, in reality, make victims more vulnerable to retaliation without legal protection. The paradoxes abound. This reality, however, is perhaps most prominent when considering the technology's potential impact on workplace culture.

Proponents of AI-based harassment monitors argue that their use can help combat "toxic workplaces" proposing that companies, "armed with this technology, . . . can protect employees, the company, and the culture from malicious employees who would otherwise be toxic to the workforce."[208] Paradoxically, use of these monitors have the potential to sow distrust, create a culture of fear, and rather than eliminate harassers, move them into other, more private but equally destructive venues.

If, as I argue above, the use of #MeTooBots exacerbates the problem of retaliation against victims of harassment by exposing them to potentially awful retaliatory actions, the result for workplace culture is problematic as well. Imagine being the coworker of an employee who was the victim of AI-detected harassment. The employee who made the harassing comments is disciplined and quietly sent to online training. In the meantime, the victim's name leaks. You look on as your coworker, who endured harassing conduct online but did not speak out about it, is removed from team projects and increasingly isolated. If management consults the legal department, they are rightly told that there is very little risk of legal liability for retaliation. Without realizing it, you absorb the impact of your coworker's treatment. You begin to understand that management is watching all the time, harassing conduct is dealt with quietly or secretly, complaints are not valued, and retaliation is tolerated or even encouraged. You then talk with other coworkers about this reality, sharing your observations and impacting their views of management. This process, in fact, creates the toxic workplace culture employers aim to combat. Rather than magically fix the organization's culture by identifying the "malicious employee," the

[205] Ajunwa, *supra* note 45, at 1673 ("The automation of decision-making processes via machine learning algorithmic systems presents itself as a legal paradox. On one hand, such automation is often an attempt to prevent unlawful discrimination, but on the other hand, there is evidence that algorithmic decision-making processes may thwart the purposes of antidiscrimination laws such as Title VII of the Civil Rights Act of 1964 and may instead serve to reproduce inequalities at scale.").
[206] Bornstein, *supra* note 45, at 523.
[207] *See generally* Bornstein, *supra* note 45; Ajunwa, *supra* note 45.
[208] See AWARE, *supra* note 3.

use of AI-based monitors allows for the slow creep of fear, distrust, and suppressed anger to infect the workplace as a whole.[209]

In addition, commentators who critique this technology as being "Orwellian" and express concern for the alleged harassers may, in fact, be highlighting another problematic impact on workplace culture. The increase in employee monitoring by #MeTooBots will likely impact employee health—both psychological and physical—and overall well-being, which can be deeply detrimental to the workplace culture. Several studies have begun to demonstrate the detrimental physical effects of workplace monitoring. For example, "[a] study by the Department of Industrial Engineering at the University of Wisconsin has shown that the introduction of intense employee monitoring at seven AT&T-owned companies led to a twenty-seven percent increase in occurrences of pain or stiffness in the shoulders, a twenty three percent increase in occurrences of neck pressure, and a twenty-one percent increase in back pain."[210] Even before the increase in digital monitoring of workers, studies demonstrated the pernicious effects of over-monitoring of a workforce. "When workers are monitored too closely, when monitoring is used to enforce difficult work pacing, and when workers feel spied on, such supervisory activities can apparently cause strong dissatisfaction. In turn, dissatisfaction with supervisory practices may negatively influence general job satisfaction and life satisfaction."[211] While companies have, for decades, put employees on notice that their IT departments are monitoring digital communications on company servers, it is quite different to understand that a machine learning system with far greater data-monitoring capabilities is reviewing all emails, chats, texts, calls, and meeting transcripts. The knowledge that monitoring on this level is not merely a risk but an actuality will likely lead to the job dissatisfaction and overall negative workplace culture results warned of in prior research.[212]

---

[209] *See generally* Eisenstadt & Geddes, *supra* note 157 (describing the impact of suppressed anger on the workplace as a whole).

[210] Eisenstadt, *Data Analytics*, *supra* note 53, at 464 n.75 (2019) (quoting Simon Head, *Big Brother Goes Digital*, N.Y. REVIEW OF BOOKS (May 24, 2018), http:// www.nybooks.com/articles/2018/05/24/big-brother-goes-digital/ (citing National Workrights Institute, Electronic Monitoring: A Poor Solution to Management Problems (2017), https://www.workrights.org/nwi_privacy_comp_monitoring_poor_solution.html) (citing Patricia Sanchez Abril et al., *Blurred Boundaries: Social Media Privacy and the Twenty-First-Century Employee*, 49 AM. BUS. L.J. 63, 69 (2012) (describing negative effects of workplace monitoring); FREDERICK S. LANE III, THE NAKED EMPLOYEE: HOW TECHNOLOGY IS COMPROMISING WORK-PLACE PRIVACY 11–16 (2003)); Maureen L. Ambrose et al., *Electronic Performance Monitoring: A Consideration of Rights*, in MANAGERIAL ETHICS: MORAL MANAGEMENT OF PEOPLE AND PROCESS 61, 69–72 (Marshall Schminke ed., 1998)).

[211] Jeffrey M. Stanton, *Traditional and Electronic Monitoring from an Organizational Justice Perspective*, 15 J. BUS. & PSYCHOL. 129, 130 (2000) (citing A.G. Bedeian & L.D. Marbert L. D., *Individual Differences in Self-Perception and The Job-Life Satisfaction Relationship*, J. OF SOC. PSYCH., 109, 111–118 (1979), T.I. Chacko, *Job And Life Satisfactions: A Causal Analysis Of Their Relationships*, ACAD. OF MGMT. J., 26, 163–169 (1983). *See also* John Chalykoff and Thomas A Kochan, *Computer-Aided Monitoring: Its Influence on Employee Job Satisfaction and Turnover,* 42:4 PERS. PSYCH. 807 (1989) (showing that, for some employees, negative effects of monitoring were inherent; for others, its negative impact could be mitigated by attention to feedback/performance appraisal processes). *See generally* Brown, *supra* note 53 (contending that workplace monitoring already threatens equity for women in the workplace).

[212] Relatedly, if the machine learning tools begin to gather sufficient data, they may begin to find digital communications that are seen as precursors to harassing conduct or comments, identifying words or phrases that are statistically correlated with later harassing comments. In effect, the bots may begin to identify workers as "harassment risks" even before any objectionable communications have been detected. They will predict harassment before it occurs. While this is a tantalizing notion, as evidenced by James de Haan's argument in favor of AI as a prophylactic, it has deeply problematic implications for workplace culture and employees' trust in the system by which they are being evaluated. *See* James de Haan, *Preventing #MeToo: Artificial Intelligence, the Law, and Prophylactics,* 38 LAW & INEQ. 69, 73 (2020) ("This paper explores a deceptively obvious solution to

Lastly, when employees begin to understand what the AI monitor is and can do, they may simply move their harassing conduct elsewhere or alter their communications to circumvent its "rules." The existence of this monitoring does not alleviate the problem of harassment. In fact, it may move it from the digital sphere where it will be automatically detected to the sometimes more damaging sphere of in-person private interaction, that will again necessitate a human report. Again here, the #MeTooBot would not fix the toxic workplace culture; it might instead simply move that toxicity from the computer back to the office.

IV. Proposals: Employer Considerations & Legal Reforms

This Article has catalogued the ways in which #MeTooBots or AI-based harassment monitors will likely impact employer liability, employee vulnerability to retaliation, victim agency, and workplace culture generally. It is a fairly grim story. Nonetheless, my intention here is not to warn off employers from ever using machine learning in the workplace. That would be a foolhardy goal since the train has likely already left the station. The ability to absorb and digest massive quantities of data and to use that data to predict and act on worker behavior is too tantalizing an innovation to ignore. My aim, instead, is to propose alternate uses of machine learning technology that takes into account legal and practical implications and, frankly, makes better, more productive use of the benefits of AI. In addition, assuming employers continue to move forward with #MeTooBots, I propose doctrinal considerations that courts should undertake given the rather dramatic impact this technology may have on the law.

A. Employer Proposals

As an initial matter, it is important to point out that the problem with #MeTooBots that this Article has discussed is *not* the ability to detect harassment in digital communications. Rather, the problems emerge from (1) overconfidence in the AI's ability to detect all harassing communications given the technology's inadequate functionality[213] and (2) the automatic reporting of problematic communications to in-house counsel or Human Resources. It is these problems that give rise to both increased liability for employers who "should have known" about harassment given the constant monitoring and decreased retaliation protection and loss of agency for victims whose identities may be reported without any purposive conduct on their parts. Given this reality, it is my proposal that employers harness the benefits of the technology while abandoning its problems.

Consider, for example, a system that uses machine learning to detect inappropriate or harassing communications but reports only the department, unit, or geographic region of the impacted workers and not the names of either alleged harasser or victim. The reporting unit would be determined on a case by case basis to ensure a large enough population to maintain anonymity of the workers involved. The tool can thus gather data on trends in the organization, analyze upticks in problematic interactions, and identify departments or regions that appear to have more problematic cultures. This data can then be used by management to allocate resources where needed, to direct trainings and other refreshers to the groups that need them, to issue expressions

human error: removing the human."). The creation of a "*Minority Report* Workplace" is not an outcome that connotes healthy, happy workers.

[213] *See* Crootof, *supra* note 116, at 243-44 (describing the problems of over-trusting in AI).

of company policy, and to appoint an ombudsperson or other worker resources to insure employees have a safe avenue for complaints. The problem is not the ability to gather massive quantities of data. Data can, if treated thoughtfully, be extremely useful in understanding workplace problems and targeting solutions. As Pauline Kim has described with reference to the use of AI to examine turnover rather than run the hiring program, "In this version of the story, data and analytic tools are used to promote, rather than undermine, workplace equality."[214]

In fact, a number of smart companies are beginning to offer this type of data collection and analysis. For example, tEquitable provides a third party ombudsperson to which employees can report complaints about all kinds of workplace problems and offers resources to either escalate the complaint appropriately or find independent solutions.[215] On the back-end, the company provides aggregated data to the contracting employer so it can see the type of complaint by department or region and so the employer can take affirmative steps to make changes where needed. tEquitable customizes so-called behavioral data by industry and company but never reveals the identities of the parties involved in the complaints.[216]

The benefits of such an AI-based detecting system are innumerable. It helps the company keep regular tabs on the workplace and identify problematic areas, while incentivizing affirmative interventions to improve workplace culture. Most importantly for our purposes, it creates none of the problems of employer liability or employee vulnerability that this paper has discussed because it maintains the anonymity of the individuals involved.[217] In addition, it does not depend on a 100% accurate detection rate. If the AI monitor misidentifies a communication as harassing when it was, in fact, benign, the worst case scenario is that already compliant workers get additional training and reminders of company policy. Similarly, if the algorithm misses a problematic interaction, it does not expose the company to additional liability. And if the company has simultaneously worked to create a culture of trust, it can expect that event to be reported and handled appropriately. This approach is all upsides.

In addition, employers might consider deploying AI in the form of a "nudge." Promoted by Richard Thaler and Cass Sunstein, a nudge is essentially a simple, inexpensive, and not overly intrusive intervention and that utilizes psychology and research on the decision-making process to promote a specific choice that is in the interest of public welfare.[218] Nudges can take numerous forms but the simplest and least intrusive nudges "supply simple information to

---

[214] Pauline Kim, *Big Data And Artificial Intelligence: New Challenges For Workplace Equality*, 57 U. LOUISVILLE L. REV. 313, 327-328 (2019) (describing beneficial uses of AI to examine data about a company's workforce to, for example, understand sources of women's underrepresentation).

[215] *See* tEQUITABLE, https://www.tequitable.com/ (last visited May 20, 2021).

[216] Telephone Interview with Lisa Gelobter, CEO and Co-Founder, tEQuitable (Feb. 10, 2020).

[217] It is worth noting that this approach may present a problem if certain departments are dominated by underrepresented minorities. As discussed above, if the algorithm is trained on too narrow a universe of resources, it may create signal and bias problems, flagging benign communications from already marginalized employees more often than others. In such a case, it would be important for human managers to be aware of this potential problem and adjust responses as a result. *See supra* text accompanying notes 121-31.

[218] *See* Todd Haugh, *Nudging Corporate Compliance*, 54 Am. Bus. L. J. 683, 690-93 (2017) ("Taken all together, the concept of nudging can best be encapsulated as follows: 'Nudges are simple interventions designed to promote desirable choices--such as compliance choices--by taking advantage of psychology … [including] a growing list of mental shortcuts, cognitive biases, and psychological quirks that subconsciously influence, and often sabotage, our decisions. Nudges are designed to either harness or neutralize these tendencies, and help us make better decisions, by subtly altering the decision-making process or the mental context in which the decision is made.'" (quoting Scott Killingsworth, Behavioral Ethics: From Nudges to Norms, BRYANCAVE.COM 1 (2017), http://ethicalsystems.org/content/behavioral-ethics-nudges-norms)).

individuals or impart reminders" and "may also be called 'deliberation nudges' because they encourage 'active, reflective decisions.'"[219] Sunstein refers to these simple nudges as "educative nudges" while Ralph Hertwig and Till Grüne-Yanoff suggest the concept of "short term boosts" and include "reminders, warnings, and information such as nutrition labels" as examples.[220] Utilizing the nudge concept, AI could be used to monitor and flag inappropriate communications and then send an automatic message to recipient, sender, or both. The recipient message might say "this email has been tagged for possible harassing content. You may want to report it. Here's how the law defines harassment, here's our company policy, and here is the process for reporting…" The message could include a link to make reporting easy and explain the process for reporting and the response the employee should expect to receive.[221] On the flip side, the AI could also be used to trigger an automatic message to the sender before the email is sent that says something along the lines of "This email has been tagged for possible harassing content. Do you want to reconsider your message?..." This idea has gained some attention in the dating app context with recent reports suggesting that Tinder is experimenting with AI monitoring messages and offering an "Are you sure you want to send?" type of alert for possible inappropriate content.[222]

Alternatively or perhaps in addition to the nudge approach, employers could deploy AI in the form of chat bots with which complaining employees can correspond. Researchers in England have begun to develop this technology, "which uses a computer-controlled robot to mimic a human interviewer, for recording and reporting workplace harassment anonymously."[223] This bot is powered by the same technology that AI-based harassment monitors use—natural language processing—to understand human language. But instead of detecting and reporting it, these chatbots interact with the victim and offer a neutral non-intimidating interface that allows victims to come forward on their own and insures an appropriate and effective response. The chatbot "guides the participant through an evidence-based cognitive interview . . . to get a high quality account of what happened."[224] It avoids the need for an initial conversation with a human who might react poorly, provide his or her own commentary, or express judgment about the victim's report. Essentially, it "bypasses concerns about trust, confidentiality and doubts being cast over harassment allegations."[225] The bot can offer an anonymous report to be sent to management, allow the reporter to decline escalation of her report, or report the user's name based on her preference. There are still some risks attendant to the use of these chat bots—they can be hacked

[219] *Id.* at 710.
[220] Ralph Hertwig and Till Grüne-Yanoff, *Nudging and Boosting*: *Steering or Empowering Good Decisions*, 12(6) PERSPECTIVES ON PSYCH. SCI. 973, 977 (2017) (differentiating between nudges and boosts but suggesting overlap between educative nudges and short term boosts). *See also* CASS SUNSTEIN, THE ETHICS OF INFLUENCE: GOVERNMENT IN THE AGE OF BEHAVIORAL SCIENCE (2016).
[221] This option still carries the risk that an employee will be encouraged to report conduct before it rises to the level of unlawful harassment, thereby exposing herself to lawful retaliation under *Clark County School District v. Breeden*, 532 U.S. 268, 270–71 (2001) (per curiam). *See supra* note 176.
[222] In the dating context, Tinder is already employing AI in this way. *See* Nicolas Rivero, *Tinder is using AI to monitor DMs and tame the creeps*, QUARTZ, May 24, 2021, https://qz.com/2011998/tinder-is-using-ai-to-monitor-dms-and-tame-the-creeps/ ("Tinder is asking its users a question we all may want to consider before dashing off a message on social media: "Are you sure you want to send?" The dating app announced . . . it will use an AI algorithm to scan private messages and compare them against texts that have been reported for inappropriate language in the past.").
[223] Shaw and Elphick, *supra* note 76.
[224] *Id.*
[225] *Id.*

as was the experience with a Microsoft-created tool,[226] or the programming can be infected with bias leading the chatbot to ask inappropriate questions. At worst, if it fails, the tool could deter future reports. But it would not come with the significant problems that AI-based harassment monitors create.

Lastly, and perhaps most importantly, it is essential that employers focus far more time and resources on the creation of workplace cultures that invite respectful dissent, encourage complaints about harassment and other forms of discrimination, take seriously and investigate properly the reports that come in, and insure that retaliation is understood and prohibited. There is no magic bullet for creating this type of workplace culture. It is far more difficult than a tech solution that promises to eliminate toxic employees, but in the end, it is the only truly effective approach. When contemplating employers' impulse to turn to tech for an answer, I find an analogy to parenting useful. If I learned that my child was being bullied in school but had not discussed it with me because she was afraid I would punish her for reporting (perhaps having witnessed her siblings suffer those consequences), would I respond by installing a camera on her clothing so that I could record and observe every moment of her day? Of course not. The obvious response is to stop punishing my children for reporting problems and to work hard to create a culture of trust and safety in the home so that they feel comfortable coming to their parents with their concerns, knowing we will hear them out, take their concerns seriously, and respond with support. The creation of healthy workplace cultures, while possibly more complicated, is also the obvious solution. There is no question that it is more difficult, takes more time, and requires consistent work. There is also no question that it is the best, most effective way to reduce the incidence of sexual harassment in the workplace. You cannot eliminate all bad actors—you can create a culture that encourages reporting, takes it seriously, and deals effectively with problems that arise. That should be the ultimate goal.

B. Legal Proposals

Although the prior sections detail objections to the use of #MeTooBots and alternative proposals for the use of AI in promoting healthy workplace cultures, I acknowledge that this technology may be the new reality despite its problems. In that scenario, I am compelled to direct some proposals to Congress and the courts as judges will invariably be faced with harassment and retaliation claims that arise from AI-detected incidents. My aim here is to suggest legislative and doctrinal changes that will continue to protect victims of harassment from retaliation despite the technology's intervention.[227]

The focus here must be on retaliation and the way in which courts understand the underlying prerequisites for protection. Focusing on "opposition conduct," Matthew Green has proposed expanding the doctrine to include instances in which the alleged victim of discrimination vents, complains, or otherwise expresses his "opposition" to her treatment even if that expression is not directed to the employer.[228] Green argues that such venting constitutes "opposing unlawful conduct" under Title VII even if she did not intend to convey her message to her manager or any other representative of her employer and should, nonetheless, be grounds for protecting the worker

[226] *Id.* (discussing Microsoft's chatbot, Tay, which was hijacked and began tweeting sexist and racist views).
[227] I am not proposing alterations to the "knew or should have known" standard for employer liability in the case of AI-based harassment monitors. If employers continue to use this technology despite its problems, the increased liability is a risk that they should endure.
[228] *See* Green *supra* note 170.

against possible retaliation.[229] While I agree with this proposal, in the age of AI-based harassment monitors, it does not go far enough.

If an algorithm can detect and report harassment without the victim's knowledge or participation, she can be exposed to retaliation for her role in the incident without having taken any affirmative steps to oppose it. Given this reality, and because of the way in which Title VII's wording approaches the problem of retaliation, there are two basic solutions. Absent Congressional intervention, courts could expand the doctrine to include "silent opposition." Such an expansion would then protect those employees who are involved in AI-detected incidents of harassment but face retaliation before they have a chance to make their opposition known to the employer. This option, although possible, is unlikely given Justice Alito's strenuous objection to this approach in his concurrence in *Crawford* and the current makeup of the Court. In addition, an expansion of the doctrine to include "silent opposition" would face significant evidentiary problems at trial. How does one demonstrate silent opposition if there is no requirement of conduct of any kind?

More useful in this area would be an amendment to the retaliation provision of Title VII (and related anti-discrimination statutes) to include "status as a victim in an incident that constitutes an unlawful employment practice"[230]—essentially an expansion of the Participation Clause to include participation as a victim in the incident itself and not merely in the claim, charge, or investigation.[231] This expansion would redirect courts' focus from protection on the basis of *conduct* (opposition or participation) to a focus on the employer's *motivation* for the retaliation.[232] It should not matter if the employer is retaliating because the employee complained or because the employee was the victim of discrimination that someone else *or something else* reported. Retaliation for one's victimhood in a discriminatory incident should be prohibited under the spirit of the law. While conduct was the focus in the past, in an AI world where human reporters (or human conduct) may be unnecessary, the law must continue to prohibit retaliation that is motivated by involvement in the underlying incident.

Conclusion

Sexual harassment and its impact on workers, organizational culture, and an entity's productivity and bottom line are enormous problems in the modern workplace. As a result, the

[229] *Id.*

[230] My preference would actually be to amend the statute to include "participation in an incident that *was reported to be* an unlawful employment practice," because it would also then eliminate the problems caused by the Objectively Reasonable Belief doctrine that limits protection to those who complain of court-determined unlawful practices and would instead allow for a good-faith standard in such cases. *See supra* notes 176, 221.

[231] *See* 42 U.S. Code § 2000e–3(a).

[232] In fact, the Supreme Court has focused more intensely on employer motivation in its understanding of disparate treatment discrimination generally. In *EEOC v. Abercrombie & Fitch Stores, Inc.,* the Court looked to the employer's motivation for refusing to hire an applicant rather than its actual knowledge of the applicant's need for a religious accommodation. 575 U.S. 768, 135 S. Ct. 2028, 2033 (2015) ("Instead, the intentional discrimination provision prohibits certain motives, regardless of the state of the actor's knowledge. Motive and knowledge are separate concepts. An employer who has actual knowledge of the need for an accommodation does not violate Title VII by refusing to hire an applicant if avoiding that accommodation is not his motive. Conversely, an employer who acts with the motive of avoiding accommodation may violate Title VII even if he has no more than an unsubstantiated suspicion that accommodation would be needed."). Here, the expansion of retaliation protection would include a focus on the employer's motivation for the retaliatory conduct despite the employee's lack of opposition conduct or the employer's lack actual knowledge of such opposition.

growing innovations in machine learning seem to offer a tantalizingly "easy" solution.[233] Unfortunately, this approach leads to a far more complicated reality with significant unintended and negative consequences.

The #MeToo Movement has spotlighted both the widespread incidence of sexual harassment and the justifiable fear of retaliation that keeps victims from reporting it.[234] In addition, the last several years have seen both an exponential increase in the adoption of AI technology applicable to the workplace and a mass move towards remote work and a digital workplace. The resulting situation explains the development of #MeTooBots, AI-based systems that aim to detect and report harassment without the need for the human reporter. The developers (and the companies that have adopted the technology) understand the significant costs of workplace harassment and attempt to bypass the problems of fear and underreporting by replacing hesitant victims with an impassive bot.

But this solution, while enticing, is a mistake for employers and employees alike. For employers, the use of AI-based harassment monitors opens the door to increased liability, turning a once employer-friendly negligence standard into something akin to strict liability. For employees, the automatic reporting of potential harassment may make the victims of such incidents legally unprotected when (as occurs in many cases), the victim is retaliated against after the harassment.[235] It also comes with ethical costs in the form of damage to victim agency, voice, and healing. Perhaps most ironically, it may come with financial and organizational costs by creating a culture of fear, distrust, and over-monitoring that can lead to employee stress, physical and mental health impacts, and ultimately, lowered productivity.

Understanding this more complicated reality, should we abandon AI-based solutions to workplace problems? It would be foolish to make such a recommendation given how entrenched machine-based learning tools are becoming and the speed of technological innovation. But it is important to recognize that use of AI is not an "easy" solution to every workplace dilemma. And there are both valuable and dangerous uses for AI in the workplace. The key is understanding the impact of machine learning solutions in order to tell the difference. For example, using AI to track workplace trends and demographic shifts or to analyze employee turnover or promotion rates can offer vital information to management that may not have been possible or as easily attainable without the technology. Similarly, using AI to detect harassment and nudge the sender and/or recipient or in the form of a chat bot to provide resources to employees or offer a neutral nonjudgmental and impassive interface with HR can be invaluable tools that benefit both parties in the employment relationship.

However, when AI is used to replace either managerial or employee functions that involve adjudication, assessment, or nuanced decision-making, the likelihood of negative, unintended consequences increases. For example, AI that can "run your virtual meetings," assessing employee emotions and responses through facial recognition software and providing both immediate feedback and a full record to review[236] has multiple potential downsides including stressful monitoring, misplaced reliance on "neutral" technology that is actually operating under hidden

[233] *See, e.g.*, de Haan, *supra* note 212 (proposing AI solutions to prevent sexual harassment).
[234] *See supra* Part II.A.
[235] *See supra* Part III.B.2.
[236] *See* Arielle Pardes, *AI Can Run Your Meetings Now,* WIRED (Nov. 24, 2020), https://www.wired.com/story/ai-can-run-work-meetings-now-headroom-clockwise/ ("Headroom's software uses emotion recognition to take the temperature of the room periodically, and to gauge how much attention participants are paying to whoever's speaking. Those metrics are displayed in a window on-screen, designed mostly to give the speaker real-time feedback that can sometimes disappear in the virtual context.").

biases, and creating a massive trove of problematic data that may be relied upon for future adverse employment decisions. Similarly, the use of AI to monitor, detect, and report potential harassment moves the technology from valuable to dangerous. Constant monitoring creates unique liability issues for employers, and automatic reporting replaces all-important human discretion and the human reporter on which harassment, discrimination, and retaliation doctrine rely.

It is, no doubt, tempting to turn to AI and its capabilities to solve the workplace's thorniest problems. But the realities of the law and human nature counsel against the rush to technology in this space. In the harassment context, the solution likely lies in the difficult work of creating a workplace culture that condemns harassment and discrimination but that welcomes and takes seriously the complaints of alleged victims. AI tools that help foster such a culture are to be lauded; those that attempt to bypass or replace this hard work should be suspect.