ESSAY

OPPORTUNITIES FOR ARTIFICIAL INTELLIGENCE IN ENVIRONMENTAL COMPLIANCE

BY
ROBERT DENNEY*

The environmental compliance process has always been a significant undertaking by the United States Environmental Protection Agency (EPA). Engrained in the process is the call on the agency to use technology to foster more complete compliance with the nation's environmental statutes. While these statutes were enacted during a "data-starved" time, engrained in them are technology-forcing mandates that have prompted exponential gains in the quality of the natural environment. Now, however, EPA faces new challenges to facilitate even greater gains during a time when the agency's resources are dwindling.

This Essay analyzes how EPA has turned to technology to facilitate environmental compliance, both today and for applications in the future. The focus of this Essay is on artificial intelligence (AI) and machine learning, though the discussion also touches on more overarching forms of technology use, such as data analytics. Part II starts with a discussion of how the current environmental compliance process plays out, noting the under-compliance problem that is a consequence of EPA balancing a decrease in resources with the agency's increasing regulatory responsibilities. Part III explains how EPA currently uses technology and big data as a means to mitigate

*J.D., 2020, University of Virginia School of Law. Robert began researching this Essay while in law school. He offers thanks to his parents for supporting him throughout law school, as well as to Professor Michael Livermore for his stimulating conversation in his LawTech course that ultimately informed this Essay. Gratitude is also owed to *Environmental Law*'s members for their hard work in preparing this Essay for publication. All errors are reserved to the author, and all views are his own.

the problems discussed in Part II. For example, the agency's Next Generation Compliance initiative involves emissions technology and electronic reporting components that serve as a benchmark for how the agency will use technology moving forward.

Part IV concludes with a discussion of novel ways EPA can use AI to facilitate compliance in the future. This Part highlights two recent studies that used machine learning to predict noncompliance risk and to identify facilities that require a certain environmental permit but are currently operating without one. Problems with these future applications are also discussed, including data accuracy issues and systematic biases that may be inherent in the data used.

I. INTRODUCTION

In 2015, the German car manufacturer Volkswagen was involved in the "Dieselgate" scandal.[1] The company had used software in its diesel-powered cars that was able to recognize when the affected vehicles were being tested for nitrogen oxide emissions.[2] Using this software, the vehicles would reduce emissions during testing and then increase emissions by up to forty times once the cars resumed normal operations on the road.[3] After researchers accidentally discovered this problem, Volkswagen was found liable, and the car company has since paid over $25 billion in fines, penalties, and settlements.[4] Other vehicle manufacturers have also been found liable for similar emissions-cheating technologies, including Daimler AG, which recently settled for $1.5 billion to resolve the claims against it.[5] Dieselgate "is one of the most infamous examples of using artificial intelligence (AI) software with malicious intent," even though the AI used in the cars was relatively simple first-

---

[1] Sören Amelang & Benjamin Wehrmann, *"Dieselgate" - A Timeline of the Car Emissions Fraud Scandal in Germany*, CLEAN ENERGY WIRE (May 25, 2020), https://perma.cc/V5L3-H9K8.

[2] DAVE REJESKI ET AL., ENV'T L. INST., WHEN SOFTWARE RULES: RULE OF LAW IN THE AGE OF ARTIFICIAL INTELLIGENCE 3 (2018), https://perma.cc/D4C6-5HXC.

[3] *Id.*

[4] Mike Curley, *Daimler, Mercedes Get OK On $1.5B DOJ Emissions Deal*, LAW360 (Mar. 9, 2021), https://perma.cc/827X-5RFF.

[5] *Id.*

generation AI technology where programmers defined the parameters in which the AI system was able to operate.[6]

In contrast to Dieselgate, the use of ToxCast by the United States Environmental Protection Agency (EPA) provides an example of AI being used to promote the public good. Traditionally, the toxicity of chemicals has been verified in animal testing, but "ToxCast applies machine-learning algorithms—specifically, linear discriminate analysis—to data on chemicals' interactions obtained from *in vitro* testing to predict their toxicities."[7] In addition to avoiding the controversy associated with testing chemicals on animals, it has been estimated that ToxCast could save the government $980,000 for every toxic chemical identified through the technology.[8] However, even though ToxCast could have monumental benefits, it is not without problems. For example, ToxCast has significant computing costs, as the program involves managing a library of thousands of chemicals.[9] The program also requires that the prospective prioritization of chemicals be tested in a way that minimizes false negatives.[10] If ToxCast incorrectly rules out a chemical as not being toxic, the consequences could be deadly.

Both Dieselgate and ToxCast demonstrate that as AI becomes more complicated, concerns will arise over how it can be used in environmental compliance, both by regulators and by regulated entities. Moreover, as AI technologies advance, they will have the power to transform the environmental compliance sector. At a general level, environmental compliance concerns the monitoring, inspection, and enforcement involved in carrying out environmental statutes such as the Clean Air Act[11] (CAA) and Clean Water Act[12] (CWA). EPA and its delegated state-run programs largely oversee environmental compliance in the United States, and the process has been plagued by a number of deficiencies since the nation's major environmental statutes were enacted almost fifty years ago.

One of these deficiencies involves the fact that environmental law became an established discipline in a "data-starved" time when computing power was limited, the use of geospatial data was in its infancy, and the causal chains involved in carcinogenic chemicals were largely unknown.[13] Despite this, the new environmental statutes embraced a compliance system that relies on data collection and new

---

6 REJESKI ET AL., *supra* note 2, at 4.

7 Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L. J. 1147, 1162 (2017).

8 *Id.* at 1162–63.

9 Eric D. Watt & Richard S. Judson, *Uncertainty Quantification in ToxCast High Throughput Screening*, PLOS One, July 2018, at 2.

10 *See* David J. Dix et al., *The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals*, 95 TOXICOLOGICAL SCIS. 5, 11 (2007) (identifying hurdles ToxCast must overcome to become a useful prioritization tool).

11 Clean Air Act, 42 U.S.C. §§ 7401–7671q (2020).

12 Federal Water Pollution Control Act, 33 U.S.C. §§ 1251–1388 (2020).

13 Gregg P. Macey, *The Architecture of Ignorance*, 2013 UTAH L. REV. 1627, 1628 (2013).

technologies, using words such as "maximum achievable control technology" and "best scientific . . . data available" to command how regulated entities will comply.[14] While this new regulatory regime was technology-forcing and prompted exponential gains in the quality of the natural environment in its first couple of decades, the regime now faces hurdles as EPA learns how to use the treasure trove of environmental compliance data it has accumulated in sophisticated ways during a time when agency resources are dwindling.[15]

This Essay explains how EPA and its associated state-run programs have turned to technology to facilitate environmental compliance, both today and in potential applications in the future. The focus of this Essay is on AI (a catchall term for machines that can replicate human capabilities, such as problem-solving) and machine learning (a type of AI where a machine is trained to make complicated predictions in a manner more efficient than could be done by humans).[16] However, this Essay also pertains to simpler or more overarching forms of technology use, such as data analytics.

Part II starts with a brief discussion of how the current environmental compliance process normally plays out, along with problems associated with this process. Namely, environmental compliance in the United States is plagued by a massive under-compliance problem during a time when EPA must balance a decrease in resources with the agency's increasing regulatory responsibilities. Part III explains how EPA currently uses innovative technologies and "big data" (i.e., large volumes of available data) as a means to mitigate the problems discussed in Part II. For example, the agency's Next Generation Compliance ("Next Gen") initiative involved innovative emissions technology, electronic reporting ("e-reporting"), and data analytics components that continue to drive the agency even after the initiative concluded in 2017. Part IV concludes with a discussion of innovative ways EPA can use the compliance data it has accumulated in AI technologies in the future. This Part highlights two recent studies that used machine learning to predict facilities' risk of noncompliance and to identify facilities that require a certain environmental permit but are currently operating without one. Problems with these potential future applications are also analyzed, including data accuracy issues and systematic biases that may be inherent in the data used.

---

[14] *See id.* at 1630 (quoting the CAA and the Endangered Species Act); 42 U.S.C. § 7412(g)(2); Endangered Species Act of 1973, 16 U.S.C. §§ 1531–1544 (2018).

[15] *See* David L. Markell & Robert L. Glicksman, *Dynamic Governance in Theory and Application, Part I*, 58 ARIZ. L. REV. 563, 594–600 (2016) (describing the challenges EPA faces with declining resources and the subsequent impact on EPA's enforcement).

[16] Kristin Burnham, *Artificial Intelligence vs. Machine Learning: What's the Difference?*, NE. U. (May 6, 2020), https://perma.cc/XR58-C5B3.

II. THE CURRENT ENVIRONMENTAL COMPLIANCE PROCESS AND ITS SHORTCOMINGS

The environmental compliance regulatory system that EPA uses is based on the deterrence model, which is a model that involves the following two policy levers: 1) "frequency of inspections to enhance the probability of detection and" 2) "magnitude of sanctions."[17] If a facility needs to emit certain pollutants into the air, ground, or into a waterbody, it applies to EPA (or a state environmental agency) for a permit. If the permit is granted, monitoring devices are installed in the facility to measure the pollution, the agency conducts regular inspections of the facility, and adjudicatory enforcement actions are taken if the facility exceeds the emissions allowed under the permit or violates other relevant permit conditions.[18] EPA uses the two policy levers of the deterrence model to try to achieve an outcome that complies with the goals of environmental statutes such as the CAA and CWA, but this process is far from perfect.

To start, the deterrence model does not map seamlessly onto environmental compliance initiatives. The model "assumes that (1) all expected benefits and costs of taking an action by the regulated entity are known and (2) collecting such information is costless."[19] In the environmental compliance arena, these assumptions do not hold. Entities regulated by environmental statutes include complex wastewater, industrial, and utility facilities that "may be uncertain about legal requirements, precluding cost-benefit analyses of possible compliance activities, or might lack the internal management capabilities necessary to undertake such evaluations."[20] Moreover, the traditional process for collecting information on these facilities is not costless—it involves considerable effort in terms of reporting and inspection requirements.[21]

The imperfections of the deterrence model show, at a high level, why effective environmental compliance is difficult to achieve. This problem can also be discerned by looking at EPA specifically and the shortcomings the agency has dealt with over the years. For one, EPA and its state-run environmental compliance programs suffer from major data gap issues resulting from the difficulty of linking pollution to environmental

---

17 ELINOR BENAMI ET AL., INNOVATIONS FOR ENVIRONMENTAL COMPLIANCE: EMERGING EVIDENCE AND OPPORTUNITIES 1 (2020), https://perma.cc/3F2J-X2BY.

18 *See generally Compliance*, U.S. ENV'T PROT. AGENCY, https://perma.cc/W3YM-3DQ6 (last visited Oct. 31, 2021) (describing EPA inspections and compliance monitoring); *Enforcement*, U.S. ENV'T PROT. AGENCY, https://perma.cc/A7BR-QKXL (last visited Oct. 31, 2021) (describing the types of enforcement actions taken by EPA).

19 BENAMI ET AL., *supra* note 17, at 3.

20 *Id.*

21 *See generally, e.g.*, RAYMOND J. KOPP & PAUL R. PORTNEY, ESTIMATING ENVIRONMENTAL COMPLIANCE COSTS FOR INDUSTRY: ENGINEERING AND ECONOMIC APPROACHES (1981), https://perma.cc/Q4A9-24JV (discussing the costs expended by industries to comply with EPA regulations and the different models used to measure this cost).

degradation from a causation perspective.[22] "Data gaps haunt every scale of regulatory interest in environmental law,"[23] and "these gaps affect problem identification, causal specification, evaluation of health and environmental impacts, valuation of harm, identification of rights, the nature of policy intervention, implementation, monitoring and enforcement, and updating and refinement."[24]

By looking at specific EPA environmental compliance programs, these data gaps are easy to see. For example, one of the most prominent permitting programs under the CAA is the new source review (NSR) process, which requires regulated entities to install pollution control equipment if they build or modify facilities in a way that would create a significant increase in emissions of a regulated pollutant.[25] Comparatively, one of the most important regulatory programs under the CWA is the section 404 permit, which "regulate[s] the discharge of dredged or fill material into waters of the United States."[26] Despite the significance of these permits, EPA does not maintain complete information on NSR permits in a centralized location, and the agency lacks a systematic framework for identifying section 404 violations.[27] These problems have been attributed to the agency's limited field presence, along with the fact that "EPA identifies violations through a passive, reactive method of relying on complaints and referrals from external sources."[28] The stack test, for instance, is the traditional approach for monitoring air pollution, and it is known to be less-than-ideal because it measures pollution at specific moments in time (rather than continuously), and it involves manipulation concerns because it relies on industry's own monitoring.[29]

In addition, central to EPA's data gap problem is the fact that the difficulty of drawing cause-and-effect linkages between pollution sources and environmental degradation is the hallmark challenge of environmental law. Even if EPA accurately identifies the amount of

---

22 *See* Markell & Glicksman, *supra* note 15, at 586–90 (describing EPA's data deficiencies and lack of centralized framework for identifying violations and coordinating with state programs).

23 Macey, *supra* note 13, at 1651.

24 Markell & Glicksman, *supra* note 15, at 586.

25 U.S. ENV'T PROT. AGENCY, FACT SHEET: NEW SOURCE REVIEW (NSR), https://perma.cc/8T3A-P8AE.

26 *Permit Program Under CWA Section 404*, U.S. ENV'T PROT. AGENCY, https://perma.cc/A645-2YZR (last visited Dec. 30, 2021).

27 *See* Markell & Glicksman, *supra* note 15, at 587–88 (noting EPA's deficiencies in maintaining data and frameworks to monitor NSR permits and Section 404 violations).

28 *Id.* at 588 (quoting EPA's Office of Inspector General); OFF. INSPECTOR GEN., U.S. EPA, REP. NO. 10-P-0009, EPA NEEDS A BETTER STRATEGY TO IDENTIFY VIOLATIONS OF SECTION 404 OF THE CLEAN WATER ACT 6 (Oct. 26, 2009), https://perma.cc/7ABT-SFG4.

29 *See* Robert L. Glicksman et al., *Technological Innovation, Data Analytics, and Environmental Enforcement*, 44 ECOLOGY L.Q. 41, 70–71 (2017) (explaining that stack tests cannot accurately represent a source's full range of emissions over all times and under all conditions); *see also* U.S. Sugar Corp. v. U.S. Env't Prot. Agency, 830 F.3d 579, 632 (D.C. Cir. 2016) (explaining problems with stack tests).

pollution reduced by an enforcement action, it is difficult to link this reduction to a corresponding benefit in terms of human health and environmental quality, even though this link is sometimes necessary to establish tort liability.[30] One specific example of this problem is seen in water pollution, where EPA has shifted its regulatory focus from technology-based mandates to ambient water-quality based effluent limitations represented through total maximum daily loads (TMDLs).[31] While technology-based mandates are easy to administer because they impose numerical caps on pipe discharges, establishing TMDLs is resource-intensive and can involve vague standards that create substantial causation problems for the dispersed pollution sources that may only be environmentally significant when cumulated.[32]

These data gap and causation problems are exacerbated by the fact that EPA is facing increasing regulatory responsibilities during a time of decreasing resources. The agency has faced persistent budget cuts; for example, the 2014 funding level for EPA was less than that provided in 1977.[33] The funding for EPA's Environmental Programs and Management, which governs enforcement actions, dropped from $2.9 billion in 2010 to $2.6 billion in 2014.[34] These budget cuts, which were a focus of the Trump administration, have translated into a decrease in staffing. Coming into the new Biden administration, the agency's workforce is 8% smaller than when President Trump took office, and the largest cuts have been in the Office of Compliance and Enforcement, which is 16% smaller than it was just a few years ago.[35] These persistent budget and staff cuts have contributed to a lack of oversight at EPA resulting in events such as the Flint water crisis.[36]

Since some of the budget and staff cuts at EPA predate the Trump administration, it is wishful thinking to believe that President Biden will be able to quickly bring the agency into a more robust form. While the recently enacted Infrastructure Investment and Jobs Act[37] does provide EPA with about $60 billion in new funding, the large majority of that new funding will go to state and tribal grants rather than to efforts that will

---

30 Markell & Glicksman, *supra* note 15, at 593–94.

31 *See id.* at 604 (discussing the problem with using TMDLs to monitor the concentration of regulated pollution in accordance with state water quality standards).

32 *See id.* (contrasting the process of using TMDL standards and technology-based mandates).

33 *See id.* at 595 ("Adjusted for inflation, the 2014 funding level was still slightly below the level provided in fiscal year 1977.").

34 *Id.*

35 Paul Gallay, *The Right to Know and the Responsibility to Act: Ensuring Environmental Compliance Through Inspection, Enforcement, and Citizen Science*, ABA SECTION OF ENV'T, ENERGY, & RES. (2019).

36 *See* U.S. ENV'T PROT. AGENCY, OFF. OF INSPECTOR GEN., MANAGEMENT WEAKNESSES DELAYED RESPONSE TO FLINT WATER CRISIS (2018) (discussing the circumstances and response to Flint's water crisis due to implementation and oversight lapses in EPA).

37 Pub. L. No. 117-58, 135 Stat. 429 (2021).

rebuild and restore the agency.[38] Moreover, on December 3, 2021, the President signed a continuing resolution passed by Congress which keeps EPA's annual appropriation at the same low level as provided under the Trump administration through February 18, 2022.[39] While the Biden administration has requested a $2 billion EPA funding increase for fiscal year 2022 (including funding to support more than 1,000 new EPA employees),[40] the fate of that request is uncertain. In fact, Republicans in the Senate have threatened to move forward with a full-year continuing resolution that would maintain EPA's funding at existing levels.[41]

Contrasting with this decline in resources is the call on EPA and similar administrative agencies to increase their regulatory responsibilities. For example, the agency has been called on to assert jurisdiction over greenhouse gas emissions under the CAA,[42] and it has faced pressure to give an expansive definition to what is considered "waters of the United States" for CWA permit purposes.[43] While the number of facilities subject to CWA permits quadrupled between 1972 and 2001, EPA has responded to this increase in responsibility largely by developing a backlog of permit requests.[44] And given that President Biden has communicated expansive goals relating to environmental protection and climate change, it can be predicted that EPA's responsibilities will continue to grow in the coming years.[45]

---

38 *See Fact Sheet: EPA & The Bipartisan Infrastructure Law*, U.S. ENV'T PROT. AGENCY (Nov. 6, 2021), https://perma.cc/7Y5K-5J3C; *see also* David Coursen, *EPA Needs an Annual Appropriation That Will Fund the Biden Environmental Agenda*, THE HILL (Dec. 23, 2021), https://perma.cc/3VAF-4TA8.

39 Further Extending Government Funding Act, Pub. L. No. 117-70 (2021); *see also* Coursen, *supra* note 38.

40 SHALANDA D. YOUNG, OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, FISCAL YEAR 2022 DISCRETIONARY FUNDING REQUEST 32 (2021); U.S. Env't Prot. Agency, FY 2022 EPA Budget in Brief 15 (May 2021), https://perma.cc/P7UC-8TN8.

41 *See* Jordain Carney, *Democrats Return with Lengthy To-Do List*, THE HILL (Jan. 3, 2022), https://perma.cc/3HT8-YEDK.

42 *See, e.g.*, Massachusetts v. U.S. Env't Prot. Agency, 549 U.S. 497, 497 (2007) ("[A] group of private organizations petitioned the Environmental Protection Agency (EPA) to begin regulating the emissions of four such gases . . . under § 202(a)(1) of the Clean Air Act."); *but see* Amy Howe, *Justices Agree to Review EPA's Authority to Regulate Greenhouse Gas Emissions*, SCOTUSBLOG (Oct. 29, 2021), https://perma.cc/XGL7-MWCY (noting that the U.S. Supreme Court will soon hear a case concerning EPA's authority to regulate greenhouse gases).

43 *See, e.g.*, Rapanos v. United States, 547 U.S. 715, 715 (2006) (describing a challenge to the federal definition of regulable waters, otherwise known as "waters of the United States").

44 Markell & Glicksman, *supra* note 15, at 602.

45 *See, e.g.*, Exec. Order No. 13990, Protecting Public Health and the Environment and Restoring Science to Tackle the Climate Crisis, 86 Fed. Reg. 7037, 7037–38 (Jan. 20, 2021) (explaining President Biden's expansive policies relating to the environment and asking EPA administrator to consider certain actions).

III. CURRENT USES OF TECHNOLOGY AND BIG DATA BY EPA

The above shortcomings of the current environmental compliance process underscore the fact that EPA "needs comprehensive, accurate, and reliable data that would allow it to better target limited resources to those regions and potential pollution problems of the greatest concern."[46] In contrast to the "data-starved" time in which the Nation's environmental statutes were written into law, recent technological advances, paired with the rise of e-reporting, are enabling EPA to find and analyze such data, though current uses by the agency are largely limited to the accumulation of data and data analytics, rather than AI.[47]

Launched in 2013, EPA's Next Gen framework was intended to bring environmental compliance into the twenty-first century by "tak[ing] advantage of new monitoring and information technology and 'us[ing] what we have learned about compliance to make it easier to comply than to violate.'"[48] The framework involved the following five interconnected components: regulation and permit design, advanced emissions/pollution detection technology, e-reporting, transparency, and innovative enforcement (e.g., through data analytics and targeting).[49] While Next Gen formally concluded in 2017, it initiated advances in monitoring technology, e-reporting, and citizen participation in data analysis that continue to shape the agency's environmental compliance programs today.

Under the Next Gen framework, the agency made significant improvements in monitoring technologies. Traditional monitoring technologies, such as those used for stack tests, were problematic because they monitored emissions only sporadically, and they also tended to involve low-quality sensors that had routine calibration problems and produced false alarms.[50] Now, however, EPA has increasingly relied on fence line monitoring, which involves placing sensors on both smokestacks and on facilities' property lines to more accurately track pollution.[51] Also, in some circumstances, the agency has begun to require continuous emissions monitoring, paired with immediate feedback technology, which allows monitoring to be performed both continuously

---

[46] Markell & Glicksman, *supra* note 15, at 589 (quoting the U.S. Government Accountability Office); U.S. Government Accountability Office, Clean Water Act: Longstanding Issues Impact EPA's and States' Enforcement Efforts: Statement of Anu K. Mittal, Natural Resources and Environment Team, GAO-10-165T, at 14 (Oct. 15, 2009), https://perma.cc/YN8N-YDHG.

[47] *See generally* DAVID FREEMAN ENGSTROM ET AL., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020), https://perma.cc/N8S3-LZYA (noting that EPA has only "experiment[ed]" with AI).

[48] Markell & Glicksman, *supra* note 15, at 610 (quoting EPA Assistant Administrator for Enforcement Cynthia Giles); Cynthia Giles, *Next Generation Compliance*, 30 ENV'T F. 22, 22 (Sept.–Oct. 2013).

[49] Markell & Glicksman, *supra* note 15, at 611; *Next Generation Compliance*, U.S. ENV'T PROT. AGENCY, https://perma.cc/CKA3-7UVW (last visited Nov. 3, 2021).

[50] Glicksman et al., *supra* note 29, at 56.

[51] *Id.* at 63, 73.

and in real time.[52] One example is a monitoring network that the Oregon Department of Environmental Quality installed on the Tillamook River that transmits water quality data to a website on a two-minute, continuous interval.[53]

While these technological advances are promising, they do not come without risks. For one, the major issues involved in implementing immediate feedback technology in monitoring include "whether legal authority exists to require its use, whether the technology exists for the compliance obligations of interest, and whether the technology is cost-effective."[54] Furthermore, while at first glance it may seem that continuous monitoring of emissions, where the results are posted onto a publicly available website, would foster greater compliance under the deterrence model, there may be limits to this assumption. While continuous monitoring may serve as a signal to facilities that they are being observed more often, it could also become "like white noise or annoyances to which one becomes acclimated or desensitized over time."[55] This risk becomes likely if EPA, because of its declining resources, is unable to institute enforcement actions even though it has more reliable evidence of noncompliance.

In addition to technological advances in monitoring, EPA is transforming its environmental compliance programs through its increasing reliance on e-reporting. In some instances, the agency has required that regulated facilities use e-reporting as a condition of their permits, and EPA has described e-reporting as "not just converting paper to electronic media. It is rather a system that guides the user through the reporting process with integrated compliance assistance and data quality checks."[56] For instance, EPA's 2015 CWA National Pollution Discharge Elimination System (NPDES) e-reporting rule is influential not just because it mandated e-reporting for NPDES permits but also because it covered nonmajor facilities.[57] Historically, EPA has focused its strained resources on the biggest and most obvious polluters, but e-reporting is allowing the agency to target nonmajor facilities in a cost-effective manner.[58] This gets at the problem of regulating pollution sources that are only environmentally significant when cumulated, which is a major dilemma EPA has dealt with for years. EPA has also worked to streamline how regulated entities self-report potential environmental violations to the agency in order to qualify for penalty mitigation under EPA's Audit

[52] *Id.* at 71–72, 77.

[53] David L. Markell & Robert L. Glicksman, *Next Generation Compliance*, NAT. RES. & ENV'T, Winter 2016, at 22, 24.

[54] David A. Hindin & Jon D. Silberman, *Designing More Effective Rules and Permits*, 7 GEO. WASH. J. ENERGY & ENV'T L. 103, 111 (2016).

[55] *Id.*

[56] Glicksman et al., *supra* note 29, at 69–70 (quoting EPA).

[57] *Id.* at 70.

[58] National Pollution Discharge Elimination System (NPDES) Electronic Reporting Rule, 80 Fed. Reg. 64,064, 64,064, 64,068 (Oct. 22, 2015).

Policy.[59] Specifically, EPA now requires facilities to self-report online through the eDisclosure system within the agency's Central Data Exchange.[60] According to the agency, reporting through eDisclosure rather than through other means (such as by paper) "saves transaction costs, increases efficiency, and provides quicker decisions."[61]

Of course, there are concerns with e-reporting. While this system has created a potential treasure trove of easily accessible data that EPA can use for data analytics purposes, it does not solve the "'fox guarding the henhouse' problem" that pervades all forms of self-reporting by regulated entities.[62] E-reporting also involves problems inherent in analyzing "semi-structured" data, which is data that includes both structured and unstructured components.[63] In e-reporting for environmental compliance, the structured data includes the tags for time and location, and the unstructured data is the emissions readings.[64] The problem with this form of data is that its lack of a fixed schema may make it difficult and costly to store, and the unstructured components make it challenging to interpret relationships between data.[65] This is certainly problematic for an agency that is strained by a decreasing budget and workforce. Moreover, certain e-reporting systems require users to input data into pre-defined check boxes that may not work well when the data involved is more complicated and nuanced than the system was designed for. For example, EPA's eDisclosure system has been described as useful for addressing minor violations, though working in eDisclosure may be frustrating when the potential violation is not straightforward.[66]

Aside from the impact e-reporting can have on both EPA and regulated entities, e-reporting also has ramifications for citizen participation. One of EPA's long-standing functions is to serve as a clearinghouse for information, and e-reporting makes the agency's environmental compliance information publicly available in ways it has never been before.[67] One specific program that the agency has

---

59 Incentives for Self-Policing: Discovery, Disclosure, Correction and Prevention of Violations, 65 Fed. Reg. 19,618, 19,618 (Apr. 11, 2000).

60 Notice of eDisclosure Portal Launch: Modernizing Implementation of EPA's Self-Policing Incentive Policies, 80 Fed. Reg. 76,476, 76,476 (Dec. 9, 2015).

61 U.S. ENV'T PROT. AGENCY, OFF. OF CIVIL ENFORCEMENT, EDISCLOSURE USER'S GUIDE 8 (2015), https://perma.cc/Q78Z-DP9Z.

62 Markell & Glicksman, *supra* note 15, at 580–81.

63 Linda K. Breggin & Judith Amsalem, *Big Data and the Environment: A Survey of Initiatives and Observations Moving Forward*, 44 ENV'T L. REP. NEWS & ANALYSIS 10984, 10985–86 (2014).

64 *Id.*

65 *See What Is Semi-Structured Data?*, GEEKSFORGEEKS, https://perma.cc/JH7X-5B27 (last updated Apr. 15, 2019) ("Due to lack of a well defined structure, it can not [be] used by computer programs easily.").

66 *See* Gaines Gwathmey et al., *EPA Simplifies Disclosure and Resolution of Certain Hazardous Chemical Reporting Violations*, PAUL WEISS, (Jan. 15, 2016) https://perma.cc/W3VA-4HKJ (noting that eDisclosure "will yield faster and simpler resolution of certain straightforward violations").

67 Glicksman et al., *supra* note 29, at 84.

implemented to engage the public in e-reporting is EPA's Enforcement and Compliance History Online (ECHO) database. ECHO provides a way for laypeople to search for facilities, investigate pollution sources, read EPA enforcement cases, create enforcement-related maps, and analyze trends in compliance and enforcement data.[68]

While ECHO has been heralded as a way to foster transparency and serve "as a 'potential resource to investors and communities,'" it has also been criticized for its data problems.[69] For instance, state environmental agencies and EPA use different vocabulary in environmental compliance, raising concerns about the completeness and accuracy of what is represented in ECHO.[70] The Maine Department of Environmental Protection, for example, has explained that the issuance of a notice of violation or the proposal of an administrative consent agreement are not recognized in ECHO as enforcement actions.[71] Nonetheless, EPA is aware of these concerns and has taken steps to mitigate ECHO's data issues, including the implementation of routine data updates and modernization of the database to make it more user-friendly.[72]

Even if the compliance data made publicly available by e-reporting is accurate, there are still concerns with its use by laypeople. For one, the data must be understandable to laypeople, and studies have suggested that when government agencies publicly release summary information rather than raw data, citizen participation and the consequent environmental performance of regulated facilities improve.[73] Moreover, the availability of this data, paired with the rise of cheap and accessible monitoring technologies, could shift the burden of monitoring from the resource-constrained EPA and onto communities.[74] While this may be a positive shift because it contributes to a "democratization" of the environmental compliance process, it could also have environmental justice ramifications for communities that lack the capacity and coordination to use technology and interpret data correctly. It could also lead to a rise in citizen suits and demands for enforcement actions, posing a coordination challenge to EPA that could undermine the initial cost savings of shifting the monitoring burden onto communities.[75]

---

68 *Enforcement and Compliance History Online*, U.S. ENV'T PROT. AGENCY, https://perma.cc/HN4J-Z437 (last visited Feb. 4, 2022).

69 Glicksman et al., *supra* note 29, at 68–69; David A. Hindin & Jon D. Silberman, *Designing More Effective Rules and Permits*, 7 GEO. WASH. J. ENERGY & ENV'T L. 103, 122 (2016).

70 *See, e.g.*, *Maine Information Relating to US EPA ECHO*, MAINE DEP. ENV'T PROT., https://perma.cc/3YU7-R7DB (describing "[t]his difference in vocabulary may result in a false impression of inactivity, or inappropriately harsh activity, when reviewing a report created with ECHO").

71 *Id.*

72 Glicksman et al., *supra* note 29, at 69.

73 BENAMI ET AL., *supra* note 17, at 4.

74 Markell & Glicksman, *supra* note 15, at 623–24.

75 *See id.* at 624 (noting skepticism toward this potential burden shift).

IV. INNOVATIVE USES OF AI FOR ENVIRONMENTAL COMPLIANCE IN THE FUTURE

In a 2020 report commissioned by the Administrative Conference of the United States, a survey of AI applications in federal agencies was undertaken.[76] The report found that 45% of the sixty-four agencies canvassed have used AI technologies, though only 12% of these applications were ranked high in sophistication.[77] This shows that the use of AI by the federal government still has a long way to go, and potential uses by EPA are no exception. The report noted that EPA has "experiment[ed]" with AI and machine learning technologies, and it has even collaborated with Stanford University to research the use of these technologies to prevent significant noncompliance, though the agency still has room to improve in the future.[78]

The use of AI to aid regulatory compliance is not foreign to other agencies, particularly those at the state and local government levels. For example, the City of Boston conducted an open tournament that awarded those who developed algorithms using Yelp reviews to predict Boston restaurant health and sanitation violations.[79] A study of the winning algorithms predicted that Boston could increase its inspection productivity by 30–50% by using these algorithms.[80] Another example is found in New York City where the Mayor's Office of Data Analytics "work[ed] with the city's fire department to use machine learning to decide where to send building inspectors."[81] The fire department uses a program with data-mining capabilities developed by Oracle to analyze approximately sixty risk factors to create lists of buildings most at risk of fire.[82]

Considering the large amount of compliance data EPA has accumulated on regulated facilities, it is not difficult to surmise that the agency may be able to use this data in AI applications in the future. Two studies coming out of Stanford University have used machine learning to aid environmental compliance, and these studies show the future of environmental compliance by EPA. The first study, titled "Machine Learning for Environmental Monitoring," used a regression forest model to predict the likelihood of a facility failing a water pollution inspection mandated under the CWA.[83] To develop the model, the study used publicly available data on facility location, industry, and inspection

---

76 ENGSTROM ET AL., *supra* note 47.

77 *Id.* at 6–7, 20.

78 *Id.* at 30, 90.

79 Edward L. Glaeser et al., *Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy*, AM. ECON. REV., May 2016, at 114.

80 *Id.*

81 Coglianese & Lehr, *supra* note 7, at 1161.

82 Brian Heaton, *New York City Fights Fire with Data*, GOV. TECH. (May 8, 2015), https://perma.cc/A9W3-V7FZ.

83 Miyuki Hino et al., *Machine Learning for Environmental Monitoring*, 1 NATURE SUSTAINABILITY 583, 583–84 (2018).

history available on ECHO and EPA's Integrated Compliance Information System.[84]

The results showed that if such a system was used to predict CWA violations, EPA could detect over seven times the expected number of violations than those detected using current practices.[85] Moreover, the study considered two important constraints: state-level differences in inspection budgets and the impact of using data derived from self-reported discharge monitoring reports.[86] The results found that even if the most rigorous budget constraints were imposed, the machine could still "double the number of violations detected through inspections."[87] Furthermore, "including the self-reported information did not uniformly improve prediction accuracy," and therefore it may be beneficial to exclude such information in order to minimize concerns about data manipulation and strategic behavior by regulated facilities.[88]

Important limitations with this machine learning study were also addressed. For one, the data used to develop the model was for inspections from 2012 to 2016, though the researchers conceded that factors predictive of inspection failure may change over time, and facilities may strategically respond to the methods used to develop risk scores to evade detection.[89] The study also noted that "an important factor to consider with all data-driven approaches is that they may codify or exacerbate existing biases and forms of discrimination."[90] In predicting risk scores for regulated facilities, the main concern is that the data used could systematically drive inspection oversight away from facilities in low-income or minority areas, and this would have important environmental justice ramifications.[91] While formulating a risk score for a facility subject to an environmental permit may not have the same procedural justice issues as those underpinning risk scores currently used for defendants during sentencing and bail proceedings, they nonetheless need to be addressed when EPA uses such prediction algorithms in the future.[92]

Another study that came out of Stanford is titled "Deep Learning to Map Concentrated Animal Feeding Operations."[93] Concentrated animal feeding operations (CAFOs) are significant facilities subject to CWA NPDES permits, yet EPA has estimated that nearly 60% of these facilities

---

[84] *Id.* at 583.
[85] *Id.* at 584.
[86] *Id.* at 583, 585.
[87] *Id.* at 583.
[88] *Id.* at 585.
[89] *Id.* at 586.
[90] *Id.*
[91] *Id.*
[92] For a criticism of bail and sentencing risk scores, see, e.g., Jason Tashea, *Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions*, ABA J. (Mar. 1, 2017), https://perma.cc/L35F-URW2.
[93] Cassandra Handan-Nader & Daniel E. Ho, *Deep Learning to Map Concentrated Animal Feeding Operations*, NATURE SUSTAINABILITY, Apr. 2019, at 298, 298 ("[N]o federal agency collects accurate and consistent data on the number, size, and location of CAFOs.").

do not hold permits.[94] Under-permitting of CAFOs has been attributed to the fact that "no federal agency collects accurate and consistent data on the number, size, and location of CAFOs," and efforts to manually comb through satellite images to detect CAFOs can take years to complete.[95]

Noting these concerns, the Stanford study used deep image-learning techniques to detect poultry CAFOs in North Carolina.[96] Specifically, the study "appl[ied] a deep convolutional neural network to high-resolution satellite images" to detect an additional 589 poultry CAFOs in the state—a 15% increase in those detected via manual methods.[97] The dataset was constructed by comparing the results of a manual census of CAFO locations to high-resolution satellite images from the United States Department of Agriculture's National Agricultural Imagery Program, and the process was substantial given that CAFO images constitute only about 0.37% of the 1,684,879 images in North Carolina.[98] While the results of this study were shown to scale over time and demonstrated that authorities could detect 95% of poultry CAFOs using less than 10% of manual resources, this deep learning technique still has challenges.[99] For one, image occlusion (where a CAFO is only partially in an image) could hinder detection, though the study tried to account for this by consolidating and re-centering image-level predictions into latitude and longitude coordinates.[100] Moreover, developing such a deep learning machine is computationally intensive and this could pose a barrier to initial investment by a resource-strained agency such as EPA.[101]

V. CONCLUSION

As one scholar put it, "EPA has long faced challenges in analyzing its data efficiently and effectively and in using the resulting analysis. Increasing volumes of data, from an increasing variety of sources, collected in an increasing variety of ways, will inevitably create both challenges and new opportunities in the agency's efforts."[102] Despite these challenges, a top EPA enforcement official has been quoted as saying that data analytics is "going to grow exponentially in the coming years . . . increase[ing] our ability to use data to find serious problems, to identify

---

[94] *Id.*

[95] *Id.* (quoting the U.S. Government Accountability Office); *Concentrated Animal Feeding Operations: EPA Needs More Information and a Clearly Defined Strategy to Protect Air and Water Quality from Pollutants of Concern*, Technical Report No. GAO-08-944 (2008), https://perma.cc/LSY6-9TE8.

[96] Handan-Nader & Ho, *supra* note 93, at 299.

[97] *Id.*

[98] *Id.* at 299–300.

[99] *Id.* at 303.

[100] *Id.* at 301.

[101] *Id.* at 302.

[102] Glicksman et al., *supra* note 29, at 58–59.

criminal activity and to help us figure out where we should be focusing our time."[103]

This hope is promising, given the significant amount of compliance data the agency has amassed by improving monitoring technologies and requiring e-reporting. Future applications of this data, notably applications that use AI and machine learning algorithms, offer additional hope to aid an agency facing an increased regulatory burden while also managing a decreased budget and workforce. When using these applications, however, it is important that EPA understand and account for problems, including systematic biases in the data used and potential manipulation by regulated entities.

[103] Juan Carlos Rodriguez, *EPA Enforcement Will Stay Tough Post-Obama, Giles Says*, LAW360 (Aug. 9, 2016), https://perma.cc/DB9M-2FNB (quoting EPA Assistant Administrator for Enforcement Cynthia Giles).