Turning Fake Data into Fake News: The A.I. Training Set as a Trojan Horse of Misinformation[*]

BILL TOMLINSON,[†] DONALD J. PATTERSON[‡] & ANDREW W. TORRANCE[§]

I. INTRODUCTION

Current news is filled with instances of the increasing benefits that artificial intelligence (A.I.) may provide to society.[1] Along with this power for good, A.I. brings with it the potential for misuse and unintended consequences.[2] One critical aspect of A.I. that demands attention is the reliance of A.I. models on training sets.[3] Training sets are vast collections of data used to

[*] We wrote this article in collaboration with ChatGPT (Mar. 23, 2023, version). We did so, in part, to investigate how scholars and AI could collaborate to produce scholarship. While that system contributed substantially to the text, we are omitting it from the author list in line with the recommendation of Springer Nature, a major scientific publisher. *See Tools Such as ChatGPT Threaten Transparent Science; Here are our Ground Rules for Their Use*, 613 NATURE 612 (2023), https://www.nature.com/articles/d41586-023-00191-1 [https://perma.cc/5VHC-ST6N] ("[N]o LLM tool will be accepted as a credited author on a research paper. That is because any attribution of authorship carries with it accountability for the work, and AI tools cannot take such responsibility."). Writing this article was, in part, an experiment in a new form of scholarly production. As a consequence, some published work by our colleagues may have been inadvertently missed by the process we describe above. We beg their indulgence for any omissions resulting from our experimental writing method. Nevertheless, we have tried to incorporate relevant references wherever we could within the parameters of our experiment. One strategy to accomplish this has been to post an early draft of our article on SSRN for anyone to review.

The authors would like to thank Amanda McElfresh for her expert research and editing. This material is based upon work supported by the National Science Foundation under Grant No. DUE-2121572.

[†] Professor of Informatics and Education at the University of California, Irvine, and a researcher in the California Institute for Telecommunications and Information Technology.

[‡] Visiting Associate Professor of Informatics Donald Bren School of Information and Computer Sciences, University of California, Irvine.

[§] Paul E. Wilson Distinguished Professor of Law and Associate Dean of Graduate and International Law, University of Kansas School of Law, and Visiting Scientist, Massachusetts Institute of Technology, Sloan School of Management.

[1] *See* Arunima Sarkar, Sirin Altiok & Şebnem Güneş Söyler, *How AI Can Help the World Fight Wildfires*, WORLD ECON. F. (May 18, 2022), https://www.weforum.org/agenda/2022/05/how-ai-can-help-the-world-fight-wildfires/ [https://perma.cc/9K2X-BV6E]; Alexander Hagerup, *AI Adoption: The 'A-Ha' Moment for Finance Leaders and How To Take Advantage of AI's Potential*, FORBES (Apr. 28, 2023, 8:45 AM), https://www.forbes.com/sites/forbestechcouncil/2023/04/28/ai-adoption-the-a-ha-moment-for-finance-leaders-and-how-to-take-advantage-of-ais-potential/?sh=4fd40785e10a [https://perma.cc/W8EN-F9DM]; *How Artificial Intelligence Is Helping Tackle Environmental Challenges*, UNITED NAT'L ENV'T PROGRAM (Nov. 7, 2022), https://www.unep.org/news-and-stories/story/how-artificial-intelligence-helping-tackle-environmental-challenges [https://perma.cc/6VKC-SE2M].

[2] *See, e.g.*, Ryan Calo, A*rtificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 410–17 (2017) (describing key questions stakeholders must consider when it comes to developing AI policy).

[3] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 680–81 (2016) ("The character of the training data can have meaningful consequences for the lessons that data mining happens to learn.").

teach algorithms how to process and interpret information.[4] Training sets are the foundation upon which A.I. models are built, and the accuracy and fairness of these sets play a crucial role in determining the outcomes produced by A.I. systems.[5]

While much attention has been given to the issue of unintentional biases and inaccuracies in training sets,[6] a more nefarious and less-discussed possibility is the deliberate subversion of these sets for malicious purposes.[7] This Article examines the potential for bad actors to exploit the vulnerability of A.I. training sets by seeding them with misleading or false information in an attempt to skew the outputs of A.I. models toward misinformation or manipulation. Drawing upon Justice Oliver Wendell Holmes Jr.'s "bad man" principle,[8] we argue that it is essential to anticipate and guard against the tactics of those who would seek to undermine the integrity of A.I. models for personal gain or malicious intent.

The potential consequences of such subversion are numerous and far-reaching. By manipulating training sets, bad actors could create revisionist histories, unjustly tarnish or enhance the reputations of individuals or organizations, or promote false ideas that serve their

[4] *See id.* at 680 ("[D]ata mining learns by example. Accordingly, what a model learns depends on the examples to which it has been exposed.").
[5] *Id.* at 683–84 ("There is an old adage in computer science: 'garbage in, garbage out.' Because data mining relies on training data as ground truth, when those inputs are themselves skewed by bias or inattention, the resulting system will produce results that are at best unreliable and at worst discriminatory."); *see also* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 96–99 (2014).
[6] *See* Barocas & Selbst, *supra* note 3, at 694–714 (discussing a framework for potential liability under Title VII for discriminatory data mining); *see also* Timnit Gebru et al., *Datasheets for Datasets*, 64 COMMC'NS ACM 86, 86 (2021).
[7] *See* Barocas & Selbst, *supra* note 3, at 692–93 (discussing ways by which "[d]ata mining could . . . breath new life into traditional forms of intentional discrimination"); *see also* Kate Crawford & Ryan Calo, *There Is a Blind Spot in AI Research*, 538 NATURE 311, 312–13 (2016) (identifying the lack of "methods to assess the sustained effects of [AI] on human populations," and proposing three tools to address that gap).
[8] *See* Holmes, The Path of the Law, Address Before the Boston University School of Law (Jan. 8, 1897), *in* 10 HARV. L. REV. 457, 459 (1897). ("If you want to know the law and nothing else, you must look at it as a bad man, who cares only for the material consequences which such knowledge enables him to predict . . . .").

interests.[9] These manipulated outputs could have significant real-world effects, ranging from increasing inequality to harming worker productivity to impeding political discourse.[10]

This Article seeks to provide a comprehensive analysis of the legal tools available to combat the deliberate subversion of A.I. training sets. Legal tools include the doctrines of fraud, nuisance, libel, slander, misappropriation, privacy, and right of publicity. We will also discuss the limitations of these tools, considering the protections afforded by the First Amendment and the need to strike a delicate balance between safeguarding the integrity of A.I. models and preserving freedom of speech.

By illuminating the potential threats posed by training set subversion and proposing legal remedies to address these challenges, this Article aims to contribute to a more secure and trustworthy A.I. ecosystem. Our ultimate goal is to ensure that the transformative potential of A.I. is harnessed for the betterment of society, rather than being exploited by malicious actors for their own nefarious purposes.

II. How Training Sets for A.I. Models Work

To understand the potential for subversion in A.I. training sets, it is essential to first comprehend how these training sets function in the development of A.I. models. At their core, training sets are vast collections of data that serve as the foundational input for teaching machine learning algorithms.[11] They provide the basis for A.I. models to learn patterns, relationships, and associations, which enable them to make predictions, recognize objects, generate text, and

[9] *See* Daron Acemoglu, *Harms of AI* 31–35 (Nat'l Bureau of Econ. Rsch., Working Paper No. 29247, 2021).
[10] *See generally id.* at 18–31.
[11] *See* Amal Joby, *What Is Training Data? How It's Used in Machine Learning*, G2 (July 30, 2021), https://learn.g2.com/training-data [perma.cc/KCX9-CXYZ] ("Training data is the initial dataset used to train machine learning algorithms. Models create and refine their rules using this data. It's a set of data samples used to fit the parameters of a machine learning model to training it by example.").

perform various other tasks.[12] The quality of an A.I. model's performance is heavily influenced by the accuracy, representativeness, and comprehensiveness of the training set it is built upon.[13]

A. Data Collection and Preparation

The process of creating a training set begins with data collection.[14] Data can be gathered from a wide array of sources, such as online databases, social media platforms, websites, user-generated content, and more.[15] The collected data often includes text, images, videos, audio, and other forms of information, depending on the desired capabilities of the A.I. model being developed.[16]

Once the data has been collected, it must be cleaned and preprocessed.[17] This step involves removing duplicate or irrelevant entries, handling missing or incomplete data, and converting the data into a suitable format for the A.I. model to process.[18] This stage is critical in ensuring the quality and reliability of the training set, as any inaccuracies or biases in the data may be propagated into the A.I. model's outputs.[19]

B. Training the A.I. Model

---

[12] *See id.*
[13] *See id.* ("High-quality data translates to accurate machine learning models. Low-quality data can significantly affect the accuracy of models . . . ."); *see also* Barocas & Selbst, *supra* note 3, at 680 ("The character of the training data can have meaningful consequences for the lessons that data mining happens to learn."); *see also* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. OF MACH. LEARNING RSCH. 1 (2018) (describing findings that "demonstrate that machine learning algorithms can discriminate based on classes like race and gender").
[14] *See* Amal Job, *Big Data Analytics: How to Make Sense of Big Data*, G2 (May 28, 2021), https://www.g2.com/articles/big-data-analytics (last visited Sept. 26, 2023) ("[D]ata analytics involves four major data preparation processes: collecting, processing, cleaning, and analyzing.").
[15] Joby, *supra* note 11 ("Raw data is gathered from multiple sources, including IoT devices, social media platforms, websites, and customer feedback.").
[16] *Id.*
[17] *See* Joby, *supra* note 14.
[18] *See* Joby, *supra* note 11 ("The data is prepared by cleaning it, accounting for missing values, removing outliers, tagging data points, and loading it into suitable places for training ML algorithms.").
[19] *See id.* (noting the importance of "quality checks" on raw data since "incorrect labels can significantly affect the model's accuracy"); *see also* Ella Wilson, *How To Remove Bias in Machine Learning Training Data*, MEDIUM (May 30, 2022), https://towardsdatascience.com/how-to-remove-bias-in-machine-learning-training-data-d54967729f88 [https://perma.cc/B5D4-ECWY] (describing the different types of biases that can exist in machine learning training data and how cleaning the data can combat biased algorithms).

Once the training set has been prepared, various machine learning algorithms can be used to train the A.I. model itself. Key types of algorithms in this domain include supervised learning, unsupervised learning, and reinforcement learning.[20] These algorithms rely on different methodologies to teach the model how to recognize patterns, make decisions, and generate outputs based on the input data.[21]

In supervised learning, for example, the model is provided with labeled input–output pairs, where the "ground-truth" or desired output is explicitly known.[22] The model learns by minimizing the difference between its predictions and the true outputs.[23] Unsupervised learning, on the other hand, involves training the model to identify underlying patterns or structures in the data without being provided explicit labels or desired outputs.[24] Reinforcement learning is another type of machine learning in which an agent learns to make decisions by taking actions in an environment to achieve a goal. The agent learns from the consequences of its actions instead of being taught.

C. Data Annotation, Crowdworkers, and Edge Cases

An essential step in the creation of training sets is data annotation, where raw data is labeled and organized in a structured format that can be used to train A.I. models.[25] This task is often carried out by human annotators, known as crowdworkers, who manually assign labels and

---

[20] *See* Batta Mahesh, *Machine Learning Algorithms - A Review*, 9 INT'L J. SCI. & RSCH. 381, 383–84 (2020) (describing the different types of machine learning algorithms).
[21] *See id.*
[22] Zhi-Hua Zhou, *A Brief Introduction to Weakly Supervised Learning*, 5 NAT'L SCI. REV. 44, 44 (2018).
[23] *See id.* at 45–46.
[24] *See* Julianna Delua, *Supervised vs. Unsupervised Learning: What's the Difference?*, IBM (Mar. 12, 2021), https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning [https://perma.cc/7399-DLLX].
[25] *See* Rayan Potter, *The Importance of High-Quality Annotated Training Data Sets in the Healthcare*, MEDIUM (June 3, 2021), https://medium.com/nerd-for-tech/the-importance-of-high-quality-annotated-training-data-sets-in-the-healthcare-f7fc37376100 [https://perma.cc/J4P8-5WQC].

categories to various data points.[26] Crowdworkers play a crucial role in ensuring the quality and accuracy of the training set by providing ground-truth labels that guide the A.I. model's learning process.[27]

During the deployment of A.I. systems, crowdworkers are also responsible for handling edge cases that the model may struggle to address due to gaps or ambiguities in the training data.[28] By providing additional annotations and content for these edge cases, crowdworkers contribute to refining and improving the training set, which can then be used to enhance the A.I. model's performance in future iterations.[29]

D. Validation and Fine Tuning

After the initial training process, the A.I. model's performance is assessed using a separate validation dataset, which has not been used during the training phase.[30] This validation dataset allows developers to evaluate the model's accuracy and generalizability to unseen data.[31] If the model's performance is found to be lacking, further fine-tuning and adjustments can be made to improve its accuracy and efficiency.[32]

Throughout this process, the quality and accuracy of the training set play a critical role in shaping the A.I. model's capabilities and outputs.[33] Crowdworkers, who are responsible for data

---

[26] *See* Michael Muller et al., *Designing Ground Truth and the Social Life of Labels*, *in* CHI '21: PROCEEDINGS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1, 2–3 (2021).
[27] *Id.* at 2.
[28] *See id.* at 3 ("[D]ata science works often have to negotiate the data and the potentially multiple meanings of the data.").
[29] *See id.* at 6 ("[I]n most cases, human knowledge remains an essential dimension in generating a high-quality, labeled training set.").
[30] *See* Joby, *supra* note 11 ("The validation dataset gives the model the first taste of unseen data.").
[31] *See id.* (explaining that validation data are used for evaluation and although the model sees this dataset occasionally, it does not learn from it).
[32] *See id.* ("In the case of ML algorithms, the training set should be periodically updated to include new information.").
[33] *See id.* ("[Q]uality training data is the most significant aspect of machine learning (and artificial intelligence) than any other.").

annotation and addressing edge cases, have a direct impact on the quality of the training set.[34] Any biases, inaccuracies, or malicious alterations that are present in the initial dataset, or introduced into the training set by crowdworkers or others, have the potential to compromise the integrity and trustworthiness of the A.I. model.[35] The ease with which an A.I. model may be compromised underscores the importance of maintaining robust processes to guide the development and deployment of training sets.[36]

III. HOW BIAS IN TRAINING SETS CAN CAUSE HARM

Bias in training sets can lead to harmful consequences across various domains, as A.I. models inherit and potentially amplify the biases present in the data they are trained on.[37] Biases can manifest in several forms, including demographic, representational, or measurement biases, among others.[38] These biases can perpetuate stereotypes, reinforce existing power structures, or undermine fairness and justice.[39] This Part explores some of the ways in which bias in A.I. training sets can result in harm.[40]

A. Discrimination and Inequality

---

[34] *See* Muller et al., *supra* note 26, at 2; *see also* Potter, *supra* note 32 ("Data Annotation is a process of identifying and mapping the desired human goal into a machine-readable form through quality training methods or data. The effectiveness is directly related to the relation with the human-defined goal and how it connects with the real model usage. Primarily, how effectively the model has been trained, keeping in the goals, and the quality of training data.").

[35] *See* Barocas & Selbst, *supra* note 3, at 683–84 ("Because data mining relies on training data as ground truth, when those inputs are themselves skewed by bias or inattention, the resulting system will produce results that are at best unreliable and at worst discriminatory."); *see generally* Muller et al., *supra* note 26.

[36] *See* Barocas & Selbst, *supra* note 10 at 717–19.

[37] *See id.* at 677 ("[D]ata mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage. Unlike more substantive forms of decision making, data mining's ill effects are often traceable to human bias, conscious or unconscious.").

[38] *See* Wilson, *supra* note 19.

[39] *See, e.g.*, Barocas & Selbst, *supra* note 3, at 698.

[40] As inspiration for combatting the harms A.I. training sets, in particular, and misinformation, in general, may cause society, Andrew W. Torrance would like to acknowledge the following works: CARL T. BERGSTROM AND JEVIN D. WEST, CALLING BULLSHIT: THE ART OF SKEPTICISM IN A DATA-DRIVEN WORLD (2020); and ORLY LOBEL, THE EQUALITY MACHINE: HARNESSING DIGITAL TECHNOLOGY FOR A BRIGHTER, MORE INCLUSIVE FUTURE (2022).

One significant issue stemming from biased training sets is the potential for discriminatory outcomes.[41] When A.I. models are trained on data that underrepresent or misrepresent certain demographic groups, they may generate biased outputs that unfairly disadvantage those groups.[42] This effect can be particularly harmful in areas such as hiring, lending, or housing, where A.I. systems are increasingly used to make decisions that directly impact people's lives.[43] For example, an A.I. model trained on a dataset predominantly featuring male job applicants may struggle to accurately assess the qualifications of female applicants, leading to unfair hiring practices.

B. Misinformation and Misrepresentation

Biased training sets can also contribute to the spread of misinformation and the misrepresentation of individuals or groups.[44] A.I. models trained on biased data may inadvertently promote stereotypes, false narratives, or misleading perspectives, distorting the public's understanding of various issues.[45] For example, an A.I.-generated news summary might disproportionately focus on crime stories involving specific racial or ethnic groups if the training

[41] Barocas & Selbst, *supra* note 3.
[42] *See id.* at 684.
[43] *See id.* at 673 (recognizing that "we live in the post-civil rights era, discrimination persists in American society and is stubbornly pervasive in employment, housing, credit, and consumer markets" and when "[a]pproached without care, data mining can reproduce existing patterns of discrimination . . . or simply reflect the widespread biases that persist in society.").
[44] *See* Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, *in* KDD '17: PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 797, 797 (2017) (noting the impact algorithmic racial bias has on determinations concerning pre-trial release or detainment); *see also* Barocas & Selbst, *supra* note 3, at 694–95, 710.
[45] Corbett-Davies et al., *supra* note 44; *see also* Nicol Turner Lee, Paul Resnick & Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies To Reduce Consumer Harms*, BROOKINGS (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ [https://perma.cc/QKE9-K73H] ("For example, automated risk assessments used by U.S. judges to determine bail and sentencing limits can generate incorrect conclusions, resulting in large cumulative effects on certain groups, like longer prison sentences or higher bails imposed on people of color.").

data overrepresent such stories, perpetuating harmful stereotypes and misconceptions.[46] Even worse, it could do so because some actor deliberately engineered an overrepresentation of news with such a biased perspective in order to achieve skewed responses from an A.I., perhaps to support a particular political ideology, to catalyze discriminatory behavior, or even to foment social unrest.[47] Another type of misinformation could include an A.I. favoring one brand of product over another due to efforts to seed a training set with biased perspectives on the two products. It is easy to imagine arms races focused on training sets arising among fierce commercial competitors hoping to engineer an edge in the marketplace for their particular products. Subverted training sets can lead to A.I.'s delivering skewed information.

C. Erosion of Trust in A.I. Systems

As instances of biased outputs from A.I. models become more widely known, public trust in these systems may be undermined.[48] The perception that A.I. models are prone to bias and may not provide reliable, unbiased results can lead to decreased adoption of potentially beneficial technologies, stymying progress and innovation in various industries.[49] Substantial caution, misunderstanding, distrust, and even fear of A.I. already exists in the minds of citizens.[50]

---

[46] "Training machines based on earlier examples can embed past prejudice and enable present-day discrimination." Eric Lander & Alondra Nelson, *Americans Need a Bill of Rights for an AI-Powered World*, WIRED (Oct. 8, 2021, 8:00 AM), https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/ [https://perma.cc/9WV3-Q6DF] (recognizing a myriad of areas where biased algorithms have the potential to perpetuate discrimination against protected classes of individuals).

[47] *See id.* ("Additionally, there's the problem of AI being deliberately abused. Some autocracies use it as a tool of state-sponsored oppression, division, and discrimination.").

[48] *See* Cynthia Dwork & Martha Minow, *Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law*, 151 DÆDALUS 309, 309 (2022) ("Social distrust of AI stems in part from incomplete and faulty data sources . . . and frequently exposed errors that reflect and amplify existing social cleavages and failures, such as racial and gender biases.").

[49] *See id.* ("[B]ig data and algorithmic tools trigger concerns over loss of control and spur decay in social trust essential for democratic governance and workable relationships in general.").

[50] *See, e.g.*, Anna Tong, *AI Threatens Humanity's Future, 61% of Americans Say: Reuters/Ipsos Poll*, REUTERS (May 17, 2023, 11:12 AM), https://www.reuters.com/technology/ai-threatens-humanitys-future-61-americans-say-reutersipsos-2023-05-17/ [https://perma.cc/B9XC-QMV4] ("More than two-thirds of Americans are concerned about the negative effects of AI and 61% believe it could threaten civilization.").

Examples of A.I.'s feeding society misinformation or bias would only exacerbate this problem, perhaps delaying or denying the arrival of the benefits of A.I.

D. Ethical and Legal Concerns

The presence of bias in training sets can also raise ethical and legal concerns, as it conflicts with the principles of fairness, accountability, and transparency that underpin responsible A.I. development.[51] Organizations employing biased A.I. systems may face legal challenges—including discrimination lawsuits—and reputational damage. As a result, companies and institutions must be proactive in addressing potential biases in their A.I. models to avoid the associated risks and liabilities.

In summary, biases in training sets can lead to a wide range of harmful consequences, from reinforcing discrimination and inequality to perpetuating misinformation and undermining trust in A.I. systems.[52] As A.I. becomes increasingly integrated into society, addressing these biases and ensuring the responsible development and deployment of A.I. models are of utmost importance to prevent harm and foster a fair and just digital ecosystem.

IV. HOW MIGHT A TRAINING SET BE SUBVERTED

Subversion of A.I. training sets can be carried out by injecting false, misleading, or biased information into the data with the intention of manipulating the model's behavior and outputs.[53] This Part discusses various ways in which training sets might be subverted, including

---

[51] *See generally* Luciano Floridi & Josh Cowls, *A Unified Framework of Five Principles for AI in Society*, HARV. DATA SCI. REV., Summer 2019, at 4–9 (proposing a unified framework for ethical AI development through a comparative analysis of recent, relevant, and reputable documents concerning ethics in AI).

[52] *See* Barocas & Selbst, *supra* note 3, at 683–84, 694–95, 710; Lander & Nelson, *supra* note 46; Dwork & Minow, *supra* note 48.

[53] *See* Will Knight, *Tainted Data Can Teach Algorithms the Wrong Lessons*, WIRED (Nov. 25, 2019, 7:00 AM), https://www.wired.com/story/tainted-data-teach-algorithms-wrong-lessons/ [https://perma.cc/6D5N-ZX6Q] ("An important leap for artificial intelligence in recent years is machines' ability to teach themselves, through endless practice, to solve problems, from mastering ancient board games to navigating busy roads. But a few subtle tweaks in the training regime can poison this 'reinforcement learning,' so that the resulting algorithm responds—like a sleeper agent—to a specified trigger by misbehaving in strange or harmful ways.").

accidental and intentional subversion, and outlines specific techniques that bad actors might employ to achieve their objectives.

A. Accidental Subversion

Accidental subversion of training sets can occur when biased or inaccurate data inadvertently find their way into the training data.[54] This can happen for several reasons:

- Flawed Data Collection Methods: Errors in data collection or sampling may lead to unrepresentative or biased data being included in the training set.[55]
- Lack of Data Quality Checks: Inadequate vetting and validation of data sources may result in the inclusion of false or misleading information in the training set.[56]
- Unintended Biases in Data Processing: Preprocessing and cleaning of data can inadvertently introduce biases, for example, by using flawed algorithms or relying on human judgments that carry inherent biases.[57]

While accidental subversion is often unintended, its impact can be just as damaging as intentional subversion, leading to biased and harmful outputs from A.I. models.

B. Intentional Subversion

---

[54] *See id.* ("AI programs can be sabotaged by the data used to train them."); *see also* MARCUS COMITER, ATTACKING ARTIFICIAL INTELLIGENCE: AI'S SECURITY VULNERABILITY AND WHAT POLICYMAKERS CAN DO ABOUT IT 13 (2019), https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf [https://perma.cc/HH4A-WQ8J] ("Because the dataset is the model's only source of knowledge, if it is corrupted or 'poisoned' by an attacker, the model learned from this data will be compromised.").

[55] *See* Barocas & Selbst, *supra* note 3, at 684 ("Decisions that depend on conclusions drawn from incorrect, partial, or nonrepresentative data may discriminate against protected classes.").

[56] *See id.* at 687–90 ("The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making."); *see also id.* at 719–20.

[57] *See id.* at 691 ("Decision makers do not necessarily intend this disparate impact because they hold prejudicial beliefs; rather, their reasonable priorities as profit seekers unintentionally recapitulate the inequality that happens to exist in society.").

In contrast to accidental subversion, intentional subversion refers to the deliberate manipulation of training sets by bad actors who aim to skew the outputs of A.I. models for malicious purposes. Intentional subversion can be carried out through various techniques:

- Data Poisoning: Bad actors inject false, misleading, or biased information into the training set with the aim of altering the model's behavior.[58] For example, they might add fake news articles, doctored images, or altered historical records to distort the model's understanding of certain events or concepts.[59]
- Adversarial Attacks: These involve exploiting vulnerabilities in the A.I. model's learning process to introduce subtle, carefully crafted perturbations in the input data, which can cause the model to produce incorrect or misleading outputs.[60] Adversarial attacks can be particularly difficult to detect and defend against, as they are often designed to be indistinguishable from genuine inputs.[61]
- Manipulation of Metadata and Labels: Bad actors may tamper with the labels or metadata associated with the training data, leading the A.I. model to learn incorrect associations or relationships.[62] For example, they might deliberately mislabel images or text to promote false narratives or to deceive the model into generating biased outputs.[63]

The potential impact of intentional subversion can be substantial, resulting in the dissemination of misinformation, the promotion of harmful ideas or ideologies, or the unjust targeting of specific individuals or groups.

---

[58] *See* Knight, *supra* note 53.
[59] *See id.*; *see also* Rowan Zellers, et al., *Defending Against Neural Fake News*, ADVANCES IN NEURAL INFO. PROCESSING SYS., no. 32, Dec. 2020.
[60] *See* COMITER, *supra* note 54, at 17–18.
[61] *See id.* at 31 ("Discovering poisoned data in order to stop poisoning attacks can be very difficult due to the scale of the datasets.").
[62] *See id.* at 30 ("[E]ven if data is collected with uncompromised equipment and stored securely, what is represented in the data itself may have been manipulated by an adversary in order to poison downstream AI systems.").
[63] *See id.*

In conclusion, subversion of training sets, whether accidental or intentional, poses a significant threat to the integrity and trustworthiness of A.I. models. To mitigate the risks associated with subversion, robust data quality controls, vigilant monitoring of training set sources, and the development of defenses against adversarial attacks are all essential. These efforts will help ensure that A.I. systems function in a fair, unbiased, and reliable manner.[64] In addition to technical defenses, the law can offer additional protections against the subversion of training sets to prevent subsequent A.I. misinformation.

C. Specific Techniques

This Section discusses the possibility that content creators, from journalists to academics to video bloggers, may be well-positioned to skew future training sets by the content they introduce into their articles, papers, or other media. Several key mechanisms can be employed to achieve this objective, which can, in turn, shape the output of future large language models (LLMs) and other A.I. systems:

- Strategic Repetition: Using certain phrases or ideas multiple times to increase the likelihood that the A.I. will recognize and adopt them as important patterns or concepts.[65]
- Framing Bias: Presenting information with a specific perspective or bias, making it more likely that the A.I. will adopt the same stance when generating content based on the training data.[66]

---

[64] *See id.* at 73–75.

[65] A key mistake, which can also be a strategic choice in some instances, could be "overtrain[ing] the model on a particular set of inputs, [so] the model becomes narrow and brittle regarding any changes that don't exactly mirror the training data." *15 Key Mistakes To Avoid When Training AI Models*, FORBES (Mar. 10, 2023, 1:15 PM), https://www.forbes.com/sites/forbestechcouncil/2023/03/10/15-key-mistakes-to-avoid-when-training-ai-models/?sh=1d022fb21ee6 [https://perma.cc/MHM4-4YHN].

[66] An AI model trained on biased data "will likely reinforce societal biases if the training data is biased." *Id.*

- Cherry-Picking Data: Selectively presenting facts or data points that support a desired narrative or conclusion, potentially leading the A.I. to develop an incomplete or skewed understanding of a topic.[67]
- Appeal to Authority: Repeatedly citing well-known figures or organizations that support a specific viewpoint to increase the credibility of that stance in the eyes of the A.I.[68]
- Misleading Analogies: Drawing comparisons between unrelated or superficially similar concepts to promote a specific belief or idea, which may confuse the A.I. or lead it to develop incorrect associations.[69]
- Use of Persuasive Language: Employing emotional or persuasive language to present certain ideas or viewpoints as more compelling or convincing, which may bias the A.I. towards these positions.[70]
- Planting False Information: Intentionally including fabricated or unverified information in the training data, which could lead the A.I. to generate content based on false premises.[71]
- Manipulating Citations: Citing sources that do not actually support the claims being made or citing non-existent sources, potentially misleading the A.I. and its users.[72]

---

[67] *See id.* (discussing how failing to use a diverse set of data can lead to biased results).
[68] *See generally* Lindsay Kramer, *Appeal to Authority Fallacy: Definition and Examples*, GRAMMARLY (Dec. 13, 2022), https://www.grammarly.com/blog/appeal-to-authority-fallacy/ [https://perma.cc/72G4-48XA] ("For an appeal to authority to be legitimate, the authority must be qualified to speak on the subject being discussed, and their statement must be directly relevant to that subject.").
[69] For a discussion on the use of analogies in AI training, see John Pavlus, *The Computer Scientist Training AI to Think with Analogies*, SCI. AM. (Aug. 6, 2021), https://www.scientificamerican.com/article/the-computer-scientist-training-ai-to-think-with-analogies/ [https://perma.cc/TDM3-TMUQ].
[70] *See id.*
[71] *See* Knight, *supra* note 53.
[72] *See generally id.*

- Crafting Self-referential Content: Creating a web of interconnected content that consistently supports a specific narrative, making it more difficult for the A.I. to identify alternative viewpoints or question the validity of the information.[73]
- Astroturfing: Generating a large volume of seemingly independent content that consistently promotes a specific viewpoint to create the illusion of widespread support or consensus, which may influence the A.I.'s perception of the topic.[74]
- Overlooking Past Work: Intentionally or accidentally excluding some scholarship or creating bias with respect to the ethnicity, nationality, religion, ideology, race, or geographic origin of an author, leading to disproportionate underweighting of contributions by such an author.[75] This could cost society dearly if it led to important contributions to knowledge being overlooked or overshadowed by later works.[76]

By acknowledging the potential risks and consequences of these techniques, we seek to render this pathway visible to A.I. developers, encouraging them to address these issues before they become widespread. In doing so, we hope to foster a more robust and fairer A.I. ecosystem that remains resilient against the malicious manipulation of training sets.

We recognize that publishing this Article represents what Oxford University Philosophy professor Nick Bostrom has termed an "information hazard," in that identifying this possible course of action may make it more likely that people will do so intentionally.[77] Nevertheless, by doing so, we seek to render this pathway visible to developers of such systems as well, and thus

---

[73] *See id.*
[74] *See* Kate Blackwood, *Lawmakers Struggle To Differentiate AI and Human Emails*, CORNELL UNIV. (Mar. 22, 2023), https://as.cornell.edu/news/lawmakers-struggle-differentiate-ai-and-human-emails [https://perma.cc/R36C-F7JB] (describing the possibility that AI astroturfing could affect the democratic process).
[75] *See 15 Key Mistakes*, *supra* note 65.
[76] *Id.*
[77] Nick Bostrom, *Information Hazards: A Typology of Potential Harms from Knowledge*, 10 REV. CONTEMP. PHIL. 44, 45 (2011) (defining "Information Hazard" as "[a] risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm.").

encourage those developers to move to address this issue before such usage becomes rampant. In addition, it is unlikely in the modern day, with so much information so freely available to so many people, that "bad men" would remain ignorant of such mischief as training data subversion simply because the phenomenon was not discussed in this Article. By raising awareness of these potential vulnerabilities, we can promote the development of proactive measures and best practices to safeguard against the manipulation of training sets.

V. LEGAL TOOLS TO FIGHT TRAINING SET SUBVERSION

In this Part, we examine several legal tools that can be employed to counteract and deter the subversion of A.I. training sets. These tools aim to protect the integrity of A.I. systems and help mitigate the adverse consequences of subverted training sets. Each legal tool addresses specific aspects of training set manipulation, offering different approaches to safeguarding A.I.-generated content and maintaining public trust in these systems.

A. Fraud

Fraud is a legal cause of action that can be directed at combating the intentional manipulation of A.I. training sets. Fraud involves a party intentionally misrepresenting facts to deceive another, leading to harm or loss.[78] In the context of A.I. training sets, fraud could arise when an individual or organization deliberately seeds false information in the data, knowing that it will be used to train A.I. models. If the A.I. then generates outputs based on this misinformation, causing harm or loss to users who rely on those outputs, a fraud claim could potentially be brought against the party responsible for the manipulation.

Several elements must be established for a successful fraud claim. These include a misrepresentation of a material fact, knowledge of the falsity, intent to deceive, justifiable

[78] RESTATEMENT (SECOND) OF TORTS § 525 (AM. L. INST. 1977).

reliance by the victim, and resulting damage.[79] In the case of A.I. training set subversion, proving these elements could be challenging, particularly in demonstrating intent and establishing a causal link between the manipulated data and the harm suffered. However, the prospect of fraud claims could serve as a sobering deterrent for those considering such actions, especially if successful cases establish legal precedents.

B. Nuisance

Nuisance is another legal doctrine that can be applied to the subversion of A.I. training sets. Nuisance generally refers to a legal cause of action in which an act is committed that unreasonably interferes with the use or enjoyment of another's property or existing rights.[80] In the context of A.I., a party manipulating training sets could be viewed as creating a "nuisance" that interferes with the proper functioning of A.I. models and the rights of users who depend on accurate A.I.-generated content.

A more traditional form of nuisance can be illustrated as follows. Imagine that Person A lives in a cabin in the woods, in part to enjoy the amenities of birdsong. Person B, who lives several kilometers away and whose activities Person A has not previously been able to overhear, purchases a stereo system complete with a giant speaker capable of broadcasting loud music that Person A can hear. Person B plays her stereo at full blast all the time, drowning out the birdsong Person A enjoys. Person A may have a cause of action against Person B for creating a nuisance with her loud stereo. If one changes "birdsong" to "accurate information," and "stereo" to

[79] *See, e.g.*, Graham v. Bank of Am., N.A., 226 Cal. App. 4th 594, 605–06 (Cal. Ct. App. 2014) (citing Perlas v. GMAC Mortg., LLC, 187 Cal. App. 4th 429, 434 (Cal. Ct. App. 2010)) ("To establish a claim for fraudulent misrepresentation, the plaintiff must prove: '(1) the defendant represented to the plaintiff that an important fact was true; (2) that representation was false; (3) the defendant knew that the representation was false when the defendant made it, or the defendant made the representation recklessly and without regard for its truth; (4) the defendant intended that the plaintiff rely on the representation; (5) the plaintiff reasonably relied on the representation; (6) the plaintiff was harmed; and (7) the plaintiff's reliance on the defendant's representation was a substantial factor in causing that harm to the plaintiff.'").

[80] RESTATEMENT (SECOND) OF TORTS § 822 (AM. L. INST. 1979).

"accurate training set subversion," one may see the analogy that might justify a nuisance (or similar) cause of action.

To establish a claim of nuisance, a plaintiff must typically prove intentional and unreasonable interference or unintentional and reckless interference; causation; and resulting harm.[81] While nuisance claims may not provide a perfect fit for addressing A.I. training set subversion, they do offer a potential avenue for seeking legal remedies. A successful claim could result in monetary damages or, perhaps even more desirably, an injunction requiring the perpetrator to cease their actions or repair damage already done.[82]

C. Libel

Libel is a form of defamation that occurs when false statements are published, causing damage to a person's reputation.[83] In the context of A.I. training set manipulation, libel could arise when false information is intentionally seeded in the data, leading A.I. models to generate defamatory content about individuals or organizations. If the false statements cause harm to the subject's reputation, a libel claim could be brought against those responsible for the manipulation.

To prove libel, a plaintiff must typically establish that the statement was false, defamatory, published, and caused harm to the plaintiff's reputation.[84] In the case of A.I. training set subversion, linking the defamatory content to the manipulation of the training set and demonstrating intent may prove more challenging than traditional models of libel. A linkage would have to be made between the subversion of the training set and harm to reputation caused

[81] *See, e.g.*, San Diego Gas & Electric Co. v. Superior Court, 13 Cal. 4th 893, 937–40 (Cal. 1996) (discussing the elements for private nuisance under California law).
[82] RESTATEMENT (SECOND) OF TORTS § 822 cmt. d (AM. L. INST. 1979) (distinguishing actions for damages from suits for injunction under a nuisance cause of action).
[83] *See* RESTATEMENT (SECOND) OF TORTS § 558 (AM. L. INST. 1977).
[84] *See id.*

by the *output* of the A.I. model so trained. However, as A.I. models become more commonplace and embedded into everyday activities, getting such a cause of action recognized, either by courts or legislatures, is likely to become easier. In any case, libel claims could still provide a deterrent effect and offer recourse to those harmed by defamatory A.I.-generated content.

D. Slander

Slander is another form of defamation, similar to libel, but involving spoken false statements that damage a person's reputation.[85] While A.I.-generated content is often written, in a literal sense, slander may become relevant in cases where A.I. models generate spoken content, such as in voice assistants or automated phone systems. In a more straightforward manner, slander could also be triggered when a human verbally repeats what they learned from a subverted A.I. If the A.I.-generated speech contains false and defamatory statements resulting from manipulated training sets, a slander claim could potentially be brought against those responsible for the subversion.

To succeed in a slander claim, a plaintiff must prove that the statement was false, defamatory, spoken, and caused harm to the plaintiff's reputation.[86] As with libel claims, establishing the elements of slander in the context of A.I. training set subversion could be challenging, particularly in demonstrating intent and causation. However, as with libel, as A.I. models become more commonplace and embedded into everyday activities, getting such a cause of action recognized, either by courts or legislatures, is likely to become easier. In addition, the prospect of slander claims may serve as another tool to deter those seeking to manipulate A.I. training sets for malicious purposes.

E. Misappropriation

[85] *Id.* § 568(2).
[86] *See id.* § 558.

Misappropriation refers to the unauthorized and unlawful use of another's property, ideas, or information for personal gain. In the context of A.I. training set manipulation, misappropriation could arise when an individual or organization deliberately seeds false information in the data, intending to exploit the A.I.-generated content for their own benefit. For example, an investor who manipulates training sets to promote their own investments could be held liable for misappropriation. In this case, the "property" misappropriated might be viewed as access to accurate information, which is *sine qua non* of many human enterprises, including journalism, investing, legal practice, medicine, and teaching.

To establish a claim for misappropriation, a plaintiff must typically prove that the defendant used the plaintiff's property or information without permission and the use was for the defendant's benefit. In the case of A.I. training set subversion, linking the manipulation to the defendant's benefit may be quite challenging. Nevertheless, as humanity relies more and more on A.I. models to provide accurate information, it is likely that successful causes of action for misappropriation will become easier to sustain. If successful, misappropriation claims could serve as a deterrent and offer legal remedies, including monetary damages and injunctions, against those who manipulate A.I. training sets for personal gain.

However, note that the Restatement of Torts has rejected misappropriation as a standalone cause of action. It can only be used in addition to another claim (e.g., trade secret misappropriation, right of publicity, breach of contract).

F. Conversion

Conversion refers to the intentional interference with the property of another.[87] This historic cause of action is typically referred to as tangible property, but some courts have expanded the definition of property to include intangible property.[88]

In the context of A.I. training set manipulation, conversion could arise when an individual or organization interferes with the information in the training set by deliberately seeding false information, intending to exploit the A.I.-generated content for their own benefit. For example, an investor who manipulates training sets to promote their own investments could be held liable for conversion.

A claim for conversion may exist when a plaintiff proves that the defendant interfered with the plaintiff's property or information without permission.[89] Establishing such interference may be a challenge in the case of A.I. training set supervision. However, as the reliance of society on A.I. models to provide accurate information grows, it will probably become easier to uphold successful causes of action for conversion. As mentioned above with regard to misappropriation, successful, conversion claims could serve as a deterrent and offer legal remedies, including monetary damages, against those who manipulate A.I. training sets for personal gain.[90]

G. Privacy

[87] RESTATEMENT (SECOND) OF TORTS § 222A (AM. L. INST. 1965).
[88] *See* Kevin G. Faley & Andrea M. Alonso, *Conversion in the Electronic Age*, MDAFP (Jan. 21, 2014), https://mdafny.com/index.aspx?TypeContent=CUSTOMPAGEARTICLE&custom_pages_articlesID=14846 [https://perma.cc/C5AR-TKBX] (discussing changes in conversion law that may allow for recognition of intangible, or electronic, property).
[89] *See* RESTATEMENT (SECOND) OF TORTS § 222A (AM. L. INST. 1965); *see also* Faley & Alonso, *supra* note 88.
[90] *See* Nick Curwen, *The Remedy in Conversion: Confusing Property and Obligation*, 26 LEGAL STUD. 570, 570 (2006).

Privacy laws protect individuals from the unauthorized use or disclosure of their personal information.[91] From humble origins,[92] the right to privacy has expanded to fill considerable legal space concerning reputation and autonomy.[93] In the context of A.I. training set subversion, privacy concerns could arise if manipulated data contains sensitive information about individuals or if the A.I.-generated content resulting from the manipulation discloses private information.[94] A privacy claim could potentially be brought against those responsible for the subversion if the manipulation leads to the violation of an individual's privacy rights.

A violation of an individual's privacy rights may lead to a privacy claim known as "publicity given to private life."[95] To succeed in this claim, a plaintiff must typically prove that their private information was disclosed and the information disclosed "would be highly offensive to a reasonable person, and . . . is not of legitimate concern to the public."[96] In the context of A.I. training set subversion, establishing causation and intent may be difficult, but privacy claims could still provide a legal remedy for those whose privacy has been violated due to manipulated A.I.-generated content.

H. Right of Publicity

Derived, in part, from the right to privacy, the right of publicity protects an individual's right to control the commercial use of their name, likeness, and other aspects of their identity.[97]

---

[91] *See generally* Kirk J. Nahra, *The Past, Present, and Future of U.S. Privacy Law*, 51 SETON HALL L. REV. 1549, 15550–54 (2021) (discussing the evolution of privacy law).
[92] Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195–96 (1890) ("[T]he question whether our law will recognize and protect the right to privacy . . . must soon come before our courts for consideration.").
[93] *See* Nahra, *supra* note 91, at 1554–63.
[94] *Id.* at 1561–62 (discussing privacy protection for sensitive data).
[95] RESTATEMENT (SECOND) OF TORTS § 652D (AM. L. INST. 1977).
[96] *Id.*
[97] RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 46 cmt. a (AM. L. INST. 1995) ("This Topic addresses the common law and statutory rules that protect the commercial value of a person's identity."). "The principal historical antecedent of the right of publicity is the right of privacy." *Id.* at cmt. b.

In the context of A.I. training set manipulation, a right of publicity claim could arise if the subversion results in the A.I.-generated content exploiting an individual's identity for commercial purposes without their consent.

To establish a right of publicity claim, a plaintiff must typically prove the unauthorized use of their identity for commercial purposes resulted in harm.[98] As with other legal claims discussed, proving causation and intent in the context of A.I. training set subversion could be challenging. Nevertheless, the right of publicity offers another avenue for legal recourse for those whose identity is exploited due to manipulated A.I.-generated content.

In conclusion, a generous array of legal tools could be employed to address the deliberate subversion of A.I. training sets. While challenges exist in establishing the necessary elements for these claims, they provide a starting point for deterring malicious actions and offering legal remedies to those harmed by manipulated A.I.-generated content. Furthermore, as society becomes increasingly reliant on A.I. models as vital sources of information, courts and legislatures are likely to become more receptive to recognizing and applying existing legal causes of action to the new class of A.I. harms. Ultimately, a comprehensive legal framework that balances free speech protections with the need to prevent and correct A.I. training set subversion will become essential for addressing this emerging threat to the prodigious benefits of A.I.[99]

VI. DATA DEFAMATION: A PROPOSED LEGAL CONCEPT BETWEEN EXISTING FRAMEWORKS AND LIMITATIONS

[98] *Id.* § 46.
[99] *See generally* Francois Candelon et al., *AI Regulation Is Coming*, HARV. BUS. REV. (Sept.–Oct. 2021), https://hbr.org/2021/09/ai-regulation-is-coming [https://perma.cc/TF76-7GZF] (discussing the need for AI regulation).

In this Article, we have explored the risks and challenges posed by subversion of A.I. training sets and have identified a range of legal tools that can be used to combat deliberate subversion. However, existing legal concepts may not fully capture the unique harms and risks posed by subversion of training data. In this Part, we propose the need for a new legal concept: data defamation. This concept would fit between the existing legal frameworks, such as fraud, nuisance, and misappropriation, and the limitations imposed by the First Amendment and other doctrines of free speech.

A. Defining Data Defamation

Data defamation refers to the deliberate introduction of false or defamatory information into a training set for the purpose of causing that A.I. to produce false or defamatory outputs.[100] Data defamation can take many forms, including the intentional inclusion of false information about individuals, companies, or institutions; the manipulation of data to promote false or misleading narratives; or the selective omission of information to promote a particular agenda. Data defamation can result in harm to individuals or groups, including reputational harm, economic harm, or other forms of harm.

B. Elements of Data Defamation

A claim of data defamation would require several elements. First, there must be an intentional introduction of false or defamatory information into a training set. This may include the deliberate manipulation of data, the use of fraudulent or misleading sources, or other forms of intentional deception. Second, the false or defamatory information must be used to train an A.I. model, and the resulting outputs must contain false or defamatory information. Third, the false

[100] This is drawn from the already existing tort of defamation found in the First Restatement of Torts, which defines defamation as "an unprivileged publication of false and defamatory matter of another which (a) is actionable irrespective of special harm, or (b) if not so actionable, is the legal cause of special harm to the other." RESTATEMENT (FIRST) OF TORTS § 558 (AM. L. INST. 1938).

or defamatory outputs must cause harm to individuals or groups, including reputational harm, economic harm, or other forms of harm.

C. Legal Precedent and Analogous Concepts

While the concept of data defamation is a novel legal concept, there are several legal precedents and analogous concepts that may be useful in developing legal approaches to combat data defamation. For example, libel and slander laws protect individuals from false or defamatory statements that harm their reputation.[101] Similarly, fraud and misrepresentation laws prohibit intentional deception and false statements that harm individuals or groups.[102] Privacy laws may also be relevant in combating data defamation, by protecting an individual's right to control the use of their personal information.[103]

VII. THE FIRST AMENDMENT AS A LIMITATION ON FIGHTING SUBVERSION

The First Amendment guarantees the freedom of speech and expression, which serves as a cornerstone of American democracy.[104] As such, any legal efforts to combat the subversion of A.I. training sets must be balanced against the protections afforded by the First Amendment. This Part will discuss the limitations imposed by the First Amendment on legal tools used to fight training set subversion and the potential justifications for overriding these protections in specific cases.

A. How the First Amendment Might Tolerate Bias in Training Sets

[101] *See supra* Parts V.C and V.D.
[102] *See supra* Part V.A.
[103] *See supra* Parts V.F and V.G.
[104] U.S. CONST. amend. I.

The First Amendment protects a wide range of speech, including ideas and opinions that may be considered biased,[105] misleading,[106] or even false.[107] Consequently, any legal efforts to regulate the content of A.I. training sets could potentially infringe upon First Amendment rights.[108] In the context of training set subversion, some instances of bias may be considered protected speech, particularly if they represent an individual's or group's opinions, beliefs, or perspectives.[109] Therefore, efforts to remove or correct such biases may not be permissible under the First Amendment.

That said, the First Amendment does not provide absolute protection for all speech.[110] As discussed in the following Section, there are certain categories of speech that receive limited or no protection, which may serve as a basis for regulating or preventing the subversion of A.I. training sets.

B. Limitations on the First Amendment

[105] For instance, the First Amendment protects much hate speech, which in some instances may be considered biased speech. *See* David Hudson, *Is Hate Speech Protected by the First Amendment?*, THE FIRE (Feb. 8, 2022), https://www.thefire.org/news/hate-speech-protected-first-amendment [https://perma.cc/9K8G-PTGK].
[106] *See* VALERIE C. BRANNON, CONG. RSCH. SERV., IF12180, FALSE SPEECH AND THE FIRST AMENDMENT: CONSTITUTIONAL LIMITS ON REGULATING MISINFORMATION 1 (2022), https://crsreports.congress.gov/product/pdf/IF/IF12180 (last visited Sept. 10, 2023).
[107] *See id.*; *see also* Eugene Volokh, *When are Lies Constitutionally Protected?*, KNIGHT FIRST AMEND. INST. (Oct. 19, 2022), https://knightcolumbia.org/content/when-are-lies-constitutionally-protected [https://perma.cc/8VRX-BWMD] (discussing the holding in New York Times Co. v. Sullivan, 376 U.S. 254 (1964), where "the Court held that even deliberate lies (said with 'actual malice') about the government are constitutionally protected," and in United States v. Alvarez, 567 U.S. 709 (2012), where "five of the justices agreed that lies about 'philosophy, religion, history, the social sciences, the arts, and the like' are generally protected.").
[108] *See* BRANNON, *supra* note 106 (recognizing the myriad areas of speech protected by the First Amendment).
[109] *See* Hudson, *supra* note 105 ("Speech that demeans on the basis of race, ethnicity, gender, religion, age, disability, or any other similar ground is hateful; but the proudest boast of our free speech jurisprudence is that we protect the freedom to express 'the thought that we hate.'").
[110] *See* Dennis v. United States, 341 U.S. 494, 508 (1951) ("Speech is not an absolute, above and beyond control by the legislature when its judgment, subject to review here, is that certain kinds of speech are so undesirable as to warrant criminal sanction. Nothing is more certain in modern society than the principle that there are no absolutes . . . .").

The First Amendment does not protect certain categories of speech, such as obscenity,[111] defamation,[112] and incitement to violence.[113] In the context of training set subversion, speech that falls within these unprotected categories could potentially be regulated without infringing upon First Amendment rights.

For instance, defamation, including libel and slander, is not protected by the First Amendment, as it involves making false and harmful statements about an individual or entity.[114] If the subversion of a training set results in A.I.-generated content that is defamatory, legal actions to remedy the harm and prevent further defamation could be pursued without violating the First Amendment.

Moreover, courts have recognized that the government may have a compelling interest in regulating certain types of speech when necessary to protect public safety, national security, or the rights of others.[115] In cases where the subversion of A.I. training sets poses a significant threat to these interests, it may be possible to argue that regulations are justified, even if they infringe on certain First Amendment protections.

In conclusion, the First Amendment imposes limitations on efforts to combat A.I. training set subversion, as it protects a broad range of speech, including biases and opinions that may be present in training data. However, certain categories of speech receive limited or no protection,

---

[111] *See* Roth v. United States, 354 U.S. 476, 485 (1957) ("We hold that obscenity is not within the area of constitutionally protected speech or press.").
[112] *See Sullivan*, 376 U.S. at 301–02 (Goldberg, J., concurring) ("The imposition of liability for private defamation does not abridge the freedom of public speech or any other freedom protected by the First Amendment.").
[113] *See* Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) ("[T]he constitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of the use of force or of law violation *except* where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action." (emphasis added)).
[114] *See id.*
[115] *See, e.g.*, First Nat'l Bank of Bos. v. Bellotti, 435 U.S. 765, 786 (1978) (citing Bates v. City of Little Rock, 361 U.S. 516, 524 (1960)) ("Especially where, as here, a prohibition is directed at speech itself, and the speech is intimately related to the process of governing, 'the State may prevail only upon showing a subordinating interest which is compelling[.]'").

and government interests in public safety, national security, or the rights of others may justify regulation in some cases. Navigating the delicate balance between First Amendment protections and the need to prevent and correct training set subversion is a complex challenge, requiring a thoughtful and nuanced legal approach.

VIII. BENEFITS TO SOCIETY OF COMBATTING TRAINING SET SUBVERSION

Preventing and addressing the subversion of A.I. training sets is crucial to maintaining the integrity of A.I. systems and ensuring their positive impact on society. In this Part, we outline some of the key benefits that arise from effectively combating training set subversion:

- Accurate and Reliable A.I.-Generated Content: Ensuring that A.I. training sets are free from intentional subversion helps produce A.I.-generated content that is more accurate, reliable, and trustworthy. This, in turn, allows individuals and organizations to make informed decisions based on the information and recommendations provided by A.I. systems, reducing the risk of negative outcomes arising from misinformation or manipulation.
- Reduced Potential for Harm: By preventing the subversion of A.I. training sets, we can mitigate the risk of A.I.-generated content causing harm to individuals or groups, either through the spread of false information, the defamation of reputations, or the promotion of biased perspectives. This contributes to a safer and more equitable digital environment for all users.
- Upholding Democratic Values: An essential aspect of a healthy democracy is the free flow of accurate information and the exchange of diverse perspectives.[116] By combatting

[116] "Democracy is built on the crucial compact that citizens will have access to reliable information and can use that information to participate in government, civic, and corporate decision-making." Eric Rosenbach & Katherine Mansted, *Can Democracy Survive in the Information Age?*, BELFER CTR. (Oct. 2018), https://www.belfercenter.org/publication/can-democracy-survive-information-age [https://perma.cc/MB7X-455Z].

training set subversion, we help ensure that A.I. systems contribute positively to the democratic process rather than promoting falsehoods or distorting public opinion through manipulation.

- Increased Public Trust in A.I.: As A.I. systems become more integrated into our daily lives, it is vital that the public has confidence in the accuracy and fairness of these systems. Addressing the issue of training set subversion is an important step in building public trust and ensuring the widespread acceptance and adoption of A.I. technologies.
- Promoting Fairness and Reducing Bias: By actively addressing the intentional subversion of A.I. training sets, we can work towards minimizing the biases that might otherwise become entrenched in A.I. systems. This helps to create more fair and equitable A.I. systems that treat all users equally and do not perpetuate harmful stereotypes or discriminatory practices.
- Enhanced Legal and Ethical Accountability: Combatting training set subversion sends a clear message that deliberate manipulation of A.I. systems for malicious purposes is unacceptable and that those responsible will be held accountable. This reinforces legal and ethical standards and helps to deter future attempts at subversion.

In light of the above, combating the subversion of A.I. training sets offers numerous benefits to society, ranging from increased accuracy and reliability of A.I.-generated content to the promotion of democratic values and the reduction of harmful biases. By addressing this issue, we can encourage, protect, and perhaps even ensure that A.I. systems continue to serve as a positive force for progress, innovation, and social good rather than deception, misinformation, and attendant harms.

IX. CONCLUSIONS

As A.I. systems continue to permeate various aspects of our lives, it is of utmost importance that we remain vigilant against the subversion of their training sets. The potential manipulation of these sets to produce biased, misleading, or even malicious A.I.-generated content poses significant risks to society, with far-reaching implications in terms of the integrity of information, fairness, and public trust.

In this Article, we have outlined the mechanisms through which training sets might be accidentally or intentionally subverted, as well as specific techniques that bad actors may employ to manipulate A.I. systems. We have also explored various legal tools available to address and combat training set subversion, including fraud, nuisance, libel, slander, misappropriation, privacy, and right of publicity, while considering the limitations imposed by the First Amendment and the need to protect freedom of speech.

The benefits of combating training set subversion are manifold. By addressing this issue, we can foster more accurate and reliable A.I.-generated content, reduce potential harm, uphold democratic values, enhance public trust, and promote fairness and accountability within A.I. systems.

As we move forward in the development and deployment of A.I. technologies, it is crucial for researchers, policymakers, and other stakeholders to work collaboratively in creating a robust framework that not only safeguards against the subversion of training sets but also promotes the ethical and responsible use of A.I.[117] By doing so, we can ensure that A.I. systems remain a powerful force for good, driving innovation and improving the quality of life for all members of society. To ignore the threat of training set subversion is to tempt the compounding of harms caused by snowballing misinformation. Justice Holmes's "bad man" must not win the

[117] *See* Floridi & Cowls, *supra* note 51.

struggle for accurate information. Society must ensure that the deep well of information from which it drinks maintains its purity, cleanliness, and clarity. Vigilance will be required not just on the proper governance of the information A.I. models generate but also on the data on which these models are trained.