

Turning Fake Data into Fake News: The AI Training Set as a Trojan Horse of Misinformation†

BILL TOMLINSON*
DONALD J. PATTERSON**
ANDREW W. TORRANCE***

† 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.

* © 2023 Bill Tomlinson. Professor of Informatics, University of California, Irvine, and Adjunct Professor of Information Management, Victoria University of Wellington–Te Herenga Waka.

** © 2023 Donald J. Patterson. Visiting Associate Professor of Informatics Donald Bren School of Information and Computer Sciences, University of California, Irvine.

*** © 2023 Andrew W. Torrance. 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.

TABLE OF CONTENTS

ABSTRACT	643
I. INTRODUCTION	644
II. HOW TRAINING SETS FOR AI MODELS WORK	646
A. <i>Data Collection and Preparation</i>	646
B. <i>Training the AI Model</i>	647
C. <i>Data Annotation, Crowdworkers, and Edge Cases</i>	648
D. <i>Validation and Fine Tuning</i>	648
III. HOW BIAS IN TRAINING SETS CAN CAUSE HARM	649
A. <i>Discrimination and Inequality</i>	650
B. <i>Misinformation and Misrepresentation</i>	650
C. <i>Erosion of Trust in AI Systems</i>	651
D. <i>Ethical and Legal Concerns</i>	652
IV. HOW MIGHT A TRAINING SET BE SUBVERTED.....	652
A. <i>Accidental Subversion</i>	653
B. <i>Intentional Subversion</i>	654
C. <i>Specific Techniques</i>	655
V. LEGAL TOOLS TO FIGHT TRAINING SET SUBVERSION	657
A. <i>Fraud</i>	657
B. <i>Nuisance</i>	658
C. <i>Libel</i>	659
D. <i>Slander</i>	660
E. <i>Misappropriation</i>	660
F. <i>Conversion</i>	661
G. <i>Privacy</i>	662
H. <i>Right of Publicity</i>	662
VI. DATA DEFAMATION: A PROPOSED LEGAL CONCEPT BETWEEN EXISTING FRAMEWORKS AND LIMITATIONS.....	663
A. <i>Defining Data Defamation</i>	664
B. <i>Elements of Data Defamation</i>	664
C. <i>Legal Precedent and Analogous Concepts</i>	664
VII. THE FIRST AMENDMENT AS A LIMITATION ON FIGHTING SUBVERSION	665
A. <i>How the First Amendment Might Tolerate Bias in Training Sets</i>	665
B. <i>Limitations on the First Amendment</i>	666
VIII. BENEFITS TO SOCIETY OF COMBATING TRAINING SET SUBVERSION	667
IX. CONCLUSIONS.....	668

ABSTRACT

Generative artificial intelligence (AI) offers tremendous benefits to society. However, these benefits must be carefully weighed against the societal damage AI can also cause. Dangers posed by inaccurate training sets have been raised by many authors. These include racial discrimination, sexual bias, and other pernicious forms of misinformation. One remedy to such problems is to ensure that training sets used to teach AI models are correct and that the data upon which they rely are accurate. An assumption behind this correction is that data inaccuracies are inadvertent mistakes. However, a darker possibility exists: the deliberate seeding of training sets with inaccurate information for the purpose of skewing the output of AI models toward misinformation. As United States Supreme Court Justice Oliver Wendell Holmes, Jr., suggested, laws are not written for the “good man,” because good people will tend to obey moral and legal principles in manners consistent with a well-functioning society even in the absence of formal laws. Rather, Justice Holmes proposed, that laws should be written with the “bad man” in mind, because bad people will push the limits of acceptable behavior, engaging in cheating, dishonesty, crime, and other societally-damaging practices, unless constrained by carefully-designed laws and their accompanying penalties.

This Article raises the spectre of the deliberate sabotage of training sets used to train AI models, with the purpose of perverting the outputs of such models. Examples include fostering revisionist histories, unjustly harming or rehabilitating the reputations of people, companies, or institutions, or even promoting as true ideas that are not. Strategic and clever efforts to introduce ideas into training sets that later manifest themselves as facts could aid and abet fraud, libel, slander, or the creation of “truth,” the belief in which promote the interests of particular individuals or groups. Imagine, for example, a first investor who buys grapefruit futures, who then seeds training sets with the idea that grapefruits will become the new gold, with the result that later prospective investors who consult AI models for investment advice are informed that they should invest in grapefruit, enriching the first investor. Or, consider a malevolent political movement that hopes to rehabilitate the reputation of an abhorrent leader; if done effectively, this movement could seed training sets with sympathetic information about this leader, resulting in positive portrayals of this leader in the future outputs of trained AI models.

This Article adopts the cautious attitude necessitated by Justice Holmes' bad man, applying it to proactively stopping, or retroactively punishing and correcting, deliberate attempts to subvert the training sets of AI models. It offers legal approaches drawn from doctrines ranging from fraud, nuisance, libel, and slander, to misappropriation, privacy, and right of publicity. It balances these with protections for speech afforded by the First Amendment and other doctrines of free speech. The result is the first comprehensive attempt to prevent, respond to, and correct deliberate attempts to subvert training sets of AI models for malicious purposes.

I. INTRODUCTION

Current news is filled with instances of the increasing benefits that artificial intelligence (AI) may provide to society.¹ Along with this power for good, AI brings with it the potential for misuse and unintended consequences.² One critical aspect of AI that demands attention is the reliance of AI models on training sets.³ Training sets are vast collections of data used to teach algorithms how to process and interpret information.⁴ Training sets are the foundation upon which AI models are built, and the accuracy and fairness of these sets play a crucial role in determining the outcomes produced by AI systems.⁵

1. See Arunima Sarkar, Sirin Altioğ & 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, *Artificial 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.”).

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).

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 AI training sets by seeding them with misleading or false information in an attempt to skew the outputs of AI 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 AI 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 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 AI 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 AI 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

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 COMM'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 . . .").

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.

to contribute to a more secure and trustworthy AI ecosystem. Our ultimate goal is to ensure that the transformative potential of AI is harnessed for the betterment of society, rather than being exploited by malicious actors for their own nefarious purposes.

II. HOW TRAINING SETS FOR AI MODELS WORK

To understand the potential for subversion in AI training sets, it is essential to first comprehend how these training sets function in the development of AI 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 AI models to learn patterns, relationships, and associations, which enable them to make predictions, recognize objects, generate text, and perform various other tasks.¹² The quality of an AI 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

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.").

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 Joby, *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.").

of information, depending on the desired capabilities of the AI 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 AI 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 AI model's outputs.¹⁹

B. Training the AI Model

Once the training set has been prepared, various machine learning algorithms can be used to train the AI 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

16. *Id.*

17. *See* Job, *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).

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.

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, Crowdfworkers, 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 AI models.²⁵ This task is often carried out by human annotators, known as crowdfworkers, who manually assign labels and categories to various data points.²⁶ Crowdfworkers play a crucial role in ensuring the quality and accuracy of the training set by providing ground-truth labels that guide the AI model's learning process.²⁷

During the deployment of AI systems, crowdfworkers 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, crowdfworkers contribute to refining and improving the training set, which can then be used to enhance the AI model's performance in future iterations.²⁹

D. Validation and Fine Tuning

After the initial training process, the AI 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

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>].

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.”).

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 AI model's capabilities and outputs.³³ Crowdworkers, who are responsible for data 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 AI model.³⁵ The ease with which an AI 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 AI 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

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.").

34. *See* Muller et al., *supra* note 26, at 2; *see also* Potter, *supra* note 25 ("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 3, 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.

undermine fairness and justice.³⁹ This Part explores some of the ways in which bias in AI training sets can result in harm.⁴⁰

A. Discrimination and Inequality

One significant issue stemming from biased training sets is the potential for discriminatory outcomes.⁴¹ When AI 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 AI systems are increasingly used to make decisions that directly impact people's lives.⁴³ For example, an AI 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 AI 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 AI-generated news summary might disproportionately

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

40. As inspiration for combatting the harms AI 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).

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,

focus on crime stories involving specific racial or ethnic groups if the training 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 AI, 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 AI 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 AI's delivering skewed information.

C. Erosion of Trust in AI Systems

As instances of biased outputs from AI models become more widely known, public trust in these systems may be undermined.⁴⁸ The perception that AI 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 AI already exists in

resulting in large cumulative effects on certain groups, like longer prison sentences or higher bails imposed on people of color.”).

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.”).

the minds of citizens.⁵⁰ Examples of AI's feeding society misinformation or bias would only exacerbate this problem, perhaps delaying or denying the arrival of the benefits of AI.

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 AI development.⁵¹ Organizations employing biased AI 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 AI 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 AI systems.⁵² As AI becomes increasingly integrated into society, addressing these biases and ensuring the responsible development and deployment of AI 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 AI 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 accidental and

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.").

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.").

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 AI models.

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.”).

B. Intentional Subversion

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 AI 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 AI 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 AI 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.

In conclusion, subversion of training sets, whether accidental or intentional, poses a significant threat to the integrity and trustworthiness of AI 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

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.*

that AI 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 AI 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 AI systems:

- Strategic Repetition: Using certain phrases or ideas multiple times to increase the likelihood that the AI 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 AI will adopt the same stance when generating content based on the training data.⁶⁶
- Cherry-Picking Data: Selectively presenting facts or data points that support a desired narrative or conclusion, potentially leading the AI 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 AI.⁶⁸

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.*

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.”).

- Misleading Analogies: Drawing comparisons between unrelated or superficially similar concepts to promote a specific belief or idea, which may confuse the AI 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 AI towards these positions.⁷⁰
- Planting False Information: Intentionally including fabricated or unverified information in the training data, which could lead the AI 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 AI and its users.⁷²
- Crafting Self-referential Content: Creating a web of interconnected content that consistently supports a specific narrative, making it more difficult for the AI 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 AI'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

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.*

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.*

By acknowledging the potential risks and consequences of these techniques, we seek to render this pathway visible to AI developers, encouraging them to address these issues before they become widespread. In doing so, we hope to foster a more robust and fairer AI 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 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 AI training sets. These tools aim to protect the integrity of AI 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 AI-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 AI training sets. Fraud involves a party intentionally misrepresenting facts to deceive another, leading to harm or loss.⁷⁸ In the context of AI training sets, fraud could arise when an

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.”).

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

individual or organization deliberately seeds false information in the data, knowing that it will be used to train AI models. If the AI 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 reliance by the victim, and resulting damage.⁷⁹ In the case of AI 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 AI 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 AI, a party manipulating training sets could be viewed as creating a "nuisance" that interferes with the proper functioning of AI models and the rights of users who depend on accurate AI-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"

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).

to “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 AI 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 AI training set manipulation, libel could arise when false information is intentionally seeded in the data, leading AI 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 AI 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 by the *output* of the AI model so trained. However, as AI 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 AI-generated content.

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.*

D. Slander

Slander is another form of defamation, similar to libel, but involving spoken false statements that damage a person's reputation.⁸⁵ While AI-generated content is often written, in a literal sense, slander may become relevant in cases where AI 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 AI. If the AI-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 AI training set subversion could be challenging, particularly in demonstrating intent and causation. However, as with libel, as AI 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 AI training sets for malicious purposes.

E. Misappropriation

Misappropriation refers to the unauthorized and unlawful use of another's property, ideas, or information for personal gain. In the context of AI training set manipulation, misappropriation could arise when an individual or organization deliberately seeds false information in the data, intending to exploit the AI-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 AI training set subversion, linking the manipulation to the defendant's benefit may be quite challenging. Nevertheless, as humanity relies more and

85. *Id.* § 568(2).

86. *See id.* § 558.

more on AI 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 AI 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 AI 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 AI-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 AI training set supervision. However, as the reliance of society on AI 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

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.

monetary damages, against those who manipulate AI training sets for personal gain.⁹⁰

G. Privacy

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 AI training set subversion, privacy concerns could arise if manipulated data contains sensitive information about individuals or if the AI-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 AI 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 AI-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 In the context of AI training set manipulation, a right of publicity claim could arise if the subversion results

90. See Nick Curwen, *The Remedy in Conversion: Confusing Property and Obligation*, 26 LEGAL STUD. 570, 570 (2006).

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 AI-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 AI 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 AI-generated content.

In conclusion, a generous array of legal tools could be employed to address the deliberate subversion of AI 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 AI-generated content. Furthermore, as society becomes increasingly reliant on AI 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 AI harms. Ultimately, a comprehensive legal framework that balances free speech protections with the need to prevent and correct AI training set subversion will become essential for addressing this emerging threat to the prodigious benefits of AI.⁹⁹

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

In this Article, we have explored the risks and challenges posed by subversion of AI 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.

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).

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 AI 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 AI model, and the resulting outputs must contain false or defamatory information. Third, the false 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

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).

101. See *supra* Parts V.0 and V.0.

102. See *supra* Part V.0.

103. See *supra* Parts V.0 and V.0.

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 AI 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*

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 AI 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.

104. U.S. CONST. amend. I.

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.’”).

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 AI training sets.

B. Limitations on the First Amendment

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 AI-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 AI 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.

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 . . .”).

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[.]’”).

In conclusion, the First Amendment imposes limitations on efforts to combat AI 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, 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 AI training sets is crucial to maintaining the integrity of AI 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 AI-Generated Content:** Ensuring that AI training sets are free from intentional subversion helps produce AI-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 AI systems, reducing the risk of negative outcomes arising from misinformation or manipulation.
- **Reduced Potential for Harm:** By preventing the subversion of AI training sets, we can mitigate the risk of AI-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 training set subversion, we help ensure that AI systems contribute positively to the

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].

democratic process rather than promoting falsehoods or distorting public opinion through manipulation.

- Increased Public Trust in AI: As AI 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 AI technologies.
- Promoting Fairness and Reducing Bias: By actively addressing the intentional subversion of AI training sets, we can work towards minimizing the biases that might otherwise become entrenched in AI systems. This helps to create more fair and equitable AI 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 AI 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 AI training sets offers numerous benefits to society, ranging from increased accuracy and reliability of AI-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 AI systems continue to serve as a positive force for progress, innovation, and social good rather than deception, misinformation, and attendant harms.

IX. CONCLUSIONS

As AI 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 AI-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 AI 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 AI-generated content, reduce potential harm, uphold democratic values, enhance public trust, and promote fairness and accountability within AI systems.

As we move forward in the development and deployment of AI 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 AI.¹¹⁷ By doing so, we can ensure that AI 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 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 AI models generate but also on the data on which these models are trained.

117. See Floridi & Cows, *supra* note 51.

