OPPORTUNITIES AND CHALLENGES OF FRONTIER DATA GOVERNANCE WITH SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data, or data generated by machine learning models, is increasingly emerging as a solution to the data access problem. However, its use introduces significant governance and accountability challenges, and potentially debases existing governance paradigms, such as compute and data governance. In this paper, we identify 3 key governance and accountability challenges that synthetic data poses - it can enable the increased emergence of malicious actors, spontaneous biases and value drift. We thus craft 3 technical mechanisms to address these specific challenges, finding applications for synthetic data towards adversarial training, bias mitigation and value reinforcement. These could not only counteract the risks of synthetic data, but serve as critical levers for governance of the frontier in the future.

1 INTRODUCTION

The rapid advancement of AI has led to an impending data bottleneck, where frontier models require exponentially increasing volumes of high-quality data, with some suggesting that the size of training corpora may exceed the sum of all human-generated data by 2030 (Villalobos et al., 2024). Empirical model capability scaling laws dictate that this scarcity bounds the overall capabilities of the frontier, regardless of strides in algorithmic complexity or computational power (Ruan et al., 2024), thus positioning it as a key issue for model developers to address. In the long term, data-efficient architectures may emerge in response to this problem and come to define the frontier. However, in the short term, it appears that synthetic data, or data generated by machine learning models as opposed to by humans, is increasingly defining the frontier. Indeed, already, a large portion of the training data of leading models in synthetic (OpenAI et al., 2024).

This shift towards synthetic data is one of many evolutions in the way models are trained that may jeopardize the efficacy of current approaches to governing and ensuring trustworthiness on the frontier. Compute governance, as proposed by Sastry et. al., for example, regulates computational power as a proxy for model capabilities (Sastry et al., 2024), but the link between the two grows more tenuous in the face of compute-efficient architectures and distilled models (DeepSeek-AI et al., 2025). Similarly, data governance, as proposed by Hausenloy et. al., relies partially on governing the flow of data through "AI Data Supply Chain" (Hausenloy et al., 2024), which they note becomes degenerate with regards to synthetic data, where the data generators, processors and trainers are the same people.

However, we propose that the advent of synthetic data, instead of being a limit to governance efforts, offers unique opportunities as a regulatory lever in addition to its challenges. Indeed, these may serve to be critical vectors for model control in the future, as existing approaches grow less effective, as detailed above.

Specifically, we identify 3 key challenges that synthetic data may pose to governance and accountability initiatives, and craft technical mechanisms to not only counter them, but establish synthetic data as a robust lever for governance of the frontier.

2 Related Work

This paper is in the style of similar proposals for governance paradigms, such as the aforementioned "compute governance" and "data governance" (Sastry et al., 2024; Hausenloy et al., 2024). Specifically, it builds upon this work by addressing a technical and temporal hole in the previous paradigms - how they will apply to synthetic data specifically, and the future more generally.

Our work lies within the subfield of Technical AI Governance (Reuel et al., 2024), as its primary application is translational: proposing technical mechanisms for policy applications, while simultaneously providing a roadmap for technical and policy work from a governance perspective.

Challenges

We outline 3 key challenges synthetic data poses to governance and accountability frameworks

2.1 Synthetic data can be used to generate misaligned data at scale.

The same ability to produce vast, tailored examples that makes synthetic data attractive to model trainers makes synthetic data attractive to malicious actors. Instead of using traditional, transparent data pipelines (Steinhardt et al., 2017), adversaries can mass-produce skewed data to deliberately misinform models (Biggio et al., 2012; Jagielski et al., 2018). Without proper safeguards, such "data poisoning" or "value hijacking" can lead to harmful ideologies or unreliable predictions in critical sectors like healthcare, finance, or public policy (Bender et al., 2021; Carlini et al., 2019).

2.2 Synthetic data can detach models from real-world contexts.

Rich synthetic environments, while useful for scalable training, risk insulating models from the dynamic value signals present in authentic human interactions. Without continual exposure to genuine linguistic subtleties, cultural norms, and ethical considerations, models may develop value systems that diverge from societal expectations (Shrivastava et al., 2017; Torralba & Efros, 2011). This insulation is further exacerbated by feedback loops where models are retrained on their own synthetic outputs, potentially entrenching misaligned values over time (Richter et al., 2020; Ganin et al., 2016).

2.3 Synthetic data could lead to spontaneous biases in black-box systems.

When large models are repeatedly retrained on their own synthetic outputs, the inherent opacity of deep learning architectures can allow small biases to accumulate unpredictably (Mehrabi et al., 2021). Over time, these biases may distort model outputs and compromise fairness, yielding results that conflict with societal expectations (Caliskan et al., 2017; Blodgett et al., 2020; Bender et al., 2021).

3 Opportunities and Mechanisms

We propose 3 mechanisms to counter the challenges outline above.

3.1 Synthetic Data for Adversarial Training

Synthetic data for adversarial training offers a scalable approach to enhance the robustness and safety of large-scale models by systematically generating malicious or deceptive scenarios. This counteracts the challenge of synthetic data for misaligned data generation. These scenarios can be used to identify and correct weaknesses in the model that might be exploited in real-world attacks. By synthetically creating adversarial examples at scale, researchers and practitioners can refine model behavior post-training, ultimately contributing to more secure frontier AI systems.

Implementation: To incorporate synthetic adversarial data in a training pipeline, one first delineates the scope of potential attacks (e.g., specific perturbations, semantic manipulations, or deceptive prompts). A generative model such as a Variational Autoencoder or a diffusion-based generator can

then be trained on a seed corpus of real-world examples, introducing adversarial constraints during data generation. This yields a large corpus of synthetic adversarial samples. These samples are integrated into the fine-tuning phase of the model, where iterative testing and updating ensure that previously uncovered weaknesses are addressed. Over multiple training rounds, the model learns to withstand a diverse set of adversarial inputs.

Existing work: Past research on adversarial training has largely relied on perturbations of real data (Goodfellow et al., 2015; Madry et al., 2018), but recent work has shown promise in generating synthetic adversarial inputs using Generative Adversarial Networks (GANs) (Creswell et al., 2018) or large-scale language models (Brown et al., 2020). Other studies have leveraged simulation frameworks and domain-specific generative models (Kingma & Welling, 2014; Dhariwal & Nichol, 2021) to produce highly varied adversarial examples that mimic real-world conditions. These approaches indicate that synthetic data can be a powerful tool in building adversarially robust systems, freeing the model developer from reliance on exhaustively labeled, human-crafted attacks.

Challenges and mitigation: While synthetic adversarial data broadens the space of potential attacks, it may also introduce novel biases if the generation process is insufficiently diverse or guided by incomplete threat models. Maintaining alignment between synthetic data distributions and real-world attack vectors can be difficult, requiring continuous monitoring and updating of generative pipelines. Additionally, the iterative feedback loop—whereby models trained on synthetic adversarial data might in turn generate subsequent synthetic data—demands careful oversight to prevent the accumulation of unrealistic or unrepresentative scenarios. Despite these challenges, synthetic adversarial data remains a valuable strategy for improving model robustness and proactively defending against the evolving landscape of security threats.

3.2 Synthetic Data for Bias Mitigation

Motivation: Real-world training datasets often suffer from demographic imbalances, such as underrepresenting certain regions or populations. For instance, health records might primarily originate from areas with highly developed healthcare systems, skewing predictive models toward those populations (Obermeyer et al., 2019). This entrenches the risk of differential treatment and can perpetuate inequities in service and care, as frontier AI systems learn more effectively from the demographics on which they have better data.

Implementation: Synthetic data can help mitigate these disparities by programmatically generating representative samples from underrepresented groups. This counteracts the challenge of spontaneous biases in synthetic data. One approach involves using generative models trained on smaller, high-quality samples from the minority population, then augmenting them with carefully designed synthetic instances (Chawla et al., 2002). In healthcare contexts, for example, generative adversarial networks have been employed to produce synthetic electronic health records that capture complex, multi-label characteristics (Choi et al., 2017). Additionally, domain experts and local stakeholders should guide the synthetic data generation process to ensure cultural and contextual fidelity.

Existing work: A growing body of literature highlights the use of generative techniques to correct or compensate for dataset biases. For example, creating synthetic faces using state-of-the-art generative adversarial networks has been shown to improve classification accuracy for underrepresented groups (Karras et al., 2019), and similar data augmentation strategies have been applied to textual data to enhance model performance (Wei & Zou, 2019). Moreover, frameworks like FairGAN have been developed to generate fairness-aware synthetic data that directly address biases in the training set (Xu et al., 2019).

Challenges and mitigation: Although synthetic data offers a promising avenue for reducing bias, it can also inadvertently introduce new biases or distort real-world distributions. Over-reliance on artificially constructed examples may yield models that perform poorly under complex, real-world conditions. Continuous monitoring, along with rigorous validation against real data, is critical. Furthermore, transparent documentation of synthetic data generation—outlining assumptions, constraints, and potential sources of error—can help stakeholders trust and verify the final models' fairness and efficacy (Mehrabi et al., 2021).

3.3 SYNTHETIC DATA FOR VALUE REINFORCEMENT

Motivation: Large-scale AI models are increasingly vulnerable to data poisoning and value hijacking, wherein adversarial actors inject harmful ideologies or manipulative content into open-source training corpora (Biggio et al., 2012; Steinhardt et al., 2017). Such attacks can distort a model's values, nudging its decisions or outputs toward harmful agendas. By contrast, synthetic data generation provides an opportunity to purposefully curate the values embedded in a training set. This counteracts the challenge of synthetic data leading to detached environments. Rather than indiscriminately scraping the web—where harmful, misleading, or biased content may dominate (Bender et al., 2021)—lab-curated synthetic corpora can emphasize collaborative, ethical, and socially constructive values.

Implementation: To implement value reinforcement via synthetic data, developers can design generative models or specialized data augmentation pipelines that focus on producing content aligned with a set of predefined principles. For instance, a language model might be guided to generate texts that uphold specific ethical frameworks or emphasize fairness and respect across different cultural perspectives (Ziegler et al., 2019). This process can include the following steps:

1. *Define Value Targets:* Collaborate with ethicists, domain experts, and stakeholders to outline desirable attributes and behaviors, translating them into clear guidelines for synthetic data generation (Amodei et al., 2016).
2. *Curated Seed Data:* Compile a smaller, high-quality corpus exemplifying the targeted values. This set serves as the seed for training or fine-tuning a generative model.
3. *Generative Pipeline:* Employ large language models, diffusion-based methods, or other generative frameworks to produce synthetic samples that faithfully reflect the curated seed's values. Mechanisms such as reinforcement learning or policy gradients can ensure alignment with these standards (Christiano et al., 2017).
4. *Validation and Iteration:* Validate generated content against established guidelines. Discard or correct any synthetic instances that deviate from the desired value set. Iteratively retrain or fine-tune the model as needed (Gehman et al., 2020).

Incentives for AI Labs: Beyond ethical considerations, AI developers have pragmatic reasons to invest in value-aligned synthetic data. Models trained on carefully curated content often demonstrate higher-quality outputs, more robust performance, and fewer public-relations liabilities. By proactively filtering out harmful or adversarial material, labs can mitigate reputational risks, reduce moderation overhead, and foster user trust. As a result, curation becomes more than a moral imperative—it is also a strategic advantage.

Challenges and mitigation: Achieving broad consensus on which values to promote can be contentious, particularly when cultural, political, or organizational perspectives diverge. Additionally, overly restrictive curation may limit the model's exposure to diverse viewpoints, potentially compromising its adaptability or realism. Regular review by multidisciplinary teams can help calibrate the balance between value alignment and open-world robustness. Finally, just as data poisoning can subvert open datasets, sophisticated attackers may attempt to introduce subtle biases into curated pipelines, necessitating continual monitoring, audits, and transparency in the curation process.

CONCLUSION

Synthetic data offers a powerful yet double-edged solution for frontier AI. It can overcome data scarcity and enhance model robustness, but without proper oversight, it risks fostering misaligned values and entrenched biases. The future of synthetic data in AI governance depends on innovative oversight mechanisms and transparent, collaborative frameworks that ensure its benefits are realized without compromising ethical standards.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Danny Mane. Concrete problems in ai safety. Technical report, OpenAI, 2016.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1467–1474, 2012.

Su Lin Blodgett, Brendan O'Connor, Benjamin Van Durme, and Lydia Green. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 529–544, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Nicholas Carlini et al. Evaluating and testing adversarial robustness in machine learning. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Edward Choi, S Biswal, Benjamin Malin, John Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete electronic health records using generative adversarial networks. In *Machine Learning for Healthcare Conference*, pp. 286–305, 2017.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4299–4307, 2017.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Bidisha Sengupta, and Anil A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 2018.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong

Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Himan Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2972–2981, 2016.

Samuel Gehman, Suchin Gururangan, Maarten Sap, and Yejin Choi. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3403–3413, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Jason Hausenloy, Duncan McClements, and Madhavendra Thakur. Towards data governance of frontier ai models, 2024. URL https://arxiv.org/abs/2412.03824.

Marcin Jagielski, Alexandru Oprea, Battista Biggio, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2819–2828, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kale Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan

Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical ai governance, 2024. URL https://arxiv.org/abs/2407.14981.

Sven Richter, Daniel Roy, and Daniel Heger. Can synthetic data bridge the reality gap in autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 100–105, 2020.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance, 2024. URL https://arxiv.org/abs/2405.10938.

Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. Computing power and the governance of artificial intelligence, 2024. URL https://arxiv.org/abs/2402.08797.

Anurag Shrivastava, Tomas Pfister, Öncel Tuzel, Joshua Susskind, Wenzhe Wang, and Rob Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2242–2251, 2017.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Antonio Torralba and Alexei C. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528, 2011.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024. URL https://arxiv.org/abs/2211.04325.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6381–6387, 2019.

Yang Xu, Yifu Zhang, Ming Liu, Haotian Zhou, and Wei Zheng. Fairgan: Fairness-aware generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 10267–10274, 2019.

Daniel M Ziegler, Nicolas Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, and Paul F Christiano. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.