Identifying and Mitigating the Security Risks of Generative AI

Other titles in Foundations and Trends® in Privacy and Security

Decentralized Finance: Protocols, Risks, and Governance
Agostino Capponi, Garud Iyengar and Jay Sethuraman
ISBN: 978-1-63828-270-9

Proofs, Arguments, and Zero-Knowledge
Justin Thaler
ISBN: 978-1-63828-124-5

Assured Autonomy Survey
Christopher Rouff and Lanier Watkins
ISBN: 978-1-63828-038-5

Hardware Platform Security for Mobile Devices
Lachlan J. Gunn, N. Asokan, Jan-Erik Ekberg, Hans Liljestrand, Vijayanand Nayani and Thomas Nyman
ISBN: 978-1-68083-976-0

Cloud Computing Security: Foundations and Research Directions
Anrin Chakraborti, Reza Curtmola, Jonathan Katz, Jason Nieh, Ahmad-Reza Sadeghi, Radu Sion and Yinqian Zhang
ISBN: 978-1-68083-958-6

Identifying and Mitigating the Security Risks of Generative AI

Clark Barrett
Brad Boyd
Elie Bursztein
Nicholas Carlini
Brad Chen
Jihye Choi
Amrita Roy Chowdhury
Mihai Christodorescu
Anupam Datta
Soheil Feizi
Kathleen Fisher
Tatsunori Hashimoto
Dan Hendrycks
Somesh Jha
Daniel Kang
Florian Kerschbaum
Eric Mitchell
John Mitchell
Zulfikar Ramzan
Khawaja Shams
Dawn Song
Ankur Taly
Diyi Yang

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Privacy and Security

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

C. Barrett *et al.*. *Identifying and Mitigating the Security Risks of Generative AI.* Foundations and Trends® in Privacy and Security, vol. 6, no. 1, pp. 1–52, 2023.

ISBN: 978-1-63828-313-3

Foundations and Trends® in Privacy and Security
Volume 6, Issue 1, 2023

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Privacy and Security publishes survey and tutorial articles in the following topics:

- Access control
- Accountability
- Anonymity
- Application security
- Artifical intelligence methods in security and privacy
- Authentication
- Big data analytics and privacy
- Cloud security
- Cyber-physical systems security and privacy
- Distributed systems security and privacy
- Embedded systems security and privacy
- Forensics
- Hardware security
- Human factors in security and privacy
- Information flow
- Intrusion detection
- Malware
- Metrics
- Mobile security and privacy
- Language-based security and privacy
- Network security
- Privacy-preserving systems
- Protocol security
- Security and privacy policies
- Security architectures
- System security
- Web security and privacy

Information for Librarians

Foundations and Trends® in Privacy and Security, 2023, Volume 6, 4 issues. ISSN paper version 2474-1558. ISSN online version 2474-1566. Also available as a combined paper and online subscription.

Contents

Identifying and Mitigating the Security Risks of Generative AI

Clark Barrett[1], Brad Boyd[1], Elie Bursztein[2], Nicholas Carlini[2], Brad Chen[2], Jihye Choi[3], Amrita Roy Chowdhury[4], Mihai Christodorescu[2], Anupam Datta[5], Soheil Feizi[6], Kathleen Fisher[7], Tatsunori Hashimoto[1], Dan Hendrycks[8], Somesh Jha[3], Daniel Kang[9], Florian Kerschbaum[10], Eric Mitchell[1], John Mitchell[1], Zulfikar Ramzan[11], Khawaja Shams[2], Dawn Song[12], Ankur Taly[2] and Diyi Yang[1]

[1] *Stanford University, USA*
[2] *Google, USA; christodorescu@google.com, kshams@google.com*
[3] *University of Wisconsin, Madison, USA; jha@cs.wisc.edu*
[4] *University of California, San Diego, USA*
[5] *Truera, USA*
[6] *University of Maryland, College Park, USA*
[7] *DARPA, USA*
[8] *Center for AI Safety, USA*
[9] *University of Illinois, Urbana Champaign, USA*
[10] *University of Waterloo, Canada*
[11] *Aura Labs, USA*
[12] *University of California, Berkeley, USA*

ABSTRACT

Every major technical invention resurfaces the dual-use dilemma—the new technology has the potential to be used

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, Kathleen Fisher, Tatsunori Hashimoto, Dan Hendrycks, Somesh Jha, Daniel Kang, Florian Kerschbaum, Eric Mitchell, John Mitchell, Zulfikar Ramzan, Khawaja Shams, Dawn Song, Ankur Taly and Diyi Yang (2023), "Identifying and Mitigating the Security Risks of Generative AI", Foundations and Trends® in Privacy and Security: Vol. 6, No. 1, pp 1–52. DOI: 10.1561/3300000041.

for good as well as for harm. Generative AI (GenAI) techniques, such as large language models (LLMs) and diffusion models, have shown remarkable capabilities (e.g., in-context learning, code-completion, and text-to-image generation and editing). However, GenAI can be used just as well by attackers to generate new attacks and increase the velocity and efficacy of existing attacks.

This monograph reports the findings of a workshop held at Google (co-organized by Stanford University and the University of Wisconsin-Madison) on the dual-use dilemma posed by GenAI. This work is not meant to be comprehensive, but is rather an attempt to synthesize some of the interesting findings from the workshop. We discuss short-term and long-term goals for the community on this topic. We hope this work provides both a launching point for a discussion on this important topic as well as interesting problems that the research community can work to address.

Keywords: robustness; behavioral, cognitive and neural learning; deep learning; security and privacy policies; security architectures; human factors in security and privacy; artificial intelligence methods in security and privacy.

1

Introduction

Emergence of powerful technologies, such as generative AI, surface the *dual-use dilemma*, which according to `Wikipedia` is defined as:

> . . . dual-use can also refer to any goods or technology which can satisfy more than one goal at any given time. Thus, expensive technologies that would otherwise benefit only civilian commercial interests can also be used to serve military purposes if they are not otherwise engaged, such as the Global Positioning System (GPS).

This dilemma was first noted with the discovery of the process for synthesizing and mass-producing ammonia which revolutionized agriculture with modern fertilizers but also led to the creation of chemical weapons during World War I. This dilemma has led to interesting policy decisions, including international treaties such as the Chemical Weapons Convention and the Treaty on the Non-Proliferation of Nuclear Weapons [97]. In computer security and cryptography, the dual-use dilemma emerges in several contexts. For example, encryption is used for protecting "data at rest," but it can also be used by ransomware to encrypt files. Similarly, anonymity techniques can help protect regular users online, but can also aid attackers to evade detection.

GenAI techniques, such as large language models (LLMs) and stable diffusion, have shown remarkable capabilities. Some of these amazing capabilities are in-context learning, code completion, and generating media that look realistic. However, GenAI has resurfaced the "dual-use dilemma," as it can be used for both productive and nefarious purposes. GenAI already provides attackers and defenders powerful access to new capabilities, and it is rapidly improving. Thus, GenAI capabilities change the landscape for malicious attacks on individuals, organizations, and a wide range of computer systems. Clumsy old "Nigerian scams" that could be detected by their primitive use of English are a thing of the past. We are also seeing the opportunity for improved defense, including monitoring of email and social media for manipulative content, as well as the potential for dramatically improved network intrusion detection, for example. Whether the rapid development and broad access to GenAI favor attackers or defenders in the long run, there are sure to be several years of unpredictability and uncertainty as the tools and our ability to use them evolve. GenAI has changed the threat landscape, and thus we need to understand it better.

To get a clearer picture of the "dual-use dilemma" for GenAI, we had a one-day workshop [4] at Google on June 27, 2023 where a group of experts convened to speak about their work. The focus of the workshop was on the following questions:

(1) How could attackers leverage GenAI technologies?

(2) How should security measures change in response to GenAI technologies?

(3) What are some current and emerging technologies we should pay attention to for designing countermeasures?

This monograph summarizes some of the findings of this workshop and puts forward several goals for both the short term and the long term.

Detailed Roadmap: Section 2 describes the capabilities of GenAI that are relevant to attacks and their defenses. Section 3 focuses on how attackers can leverage these GenAI capabilities. Section 4 investigates how defenders can leverage GenAI technologies to mitigate the risks

of these attacks. This list of attacks and defenses is not meant to be exhaustive, but it rather reflects several themes that repeatedly surfaced during the workshop. Short-term (i.e., within the next one or two years) goals for the community are discussed in Section 5. Long-term goals that correspond to challenging issues are discussed in Section 6. We end the monograph with some concluding remarks (Section 7). We acknowledge that this work is not the final word on this topic and reiterate that it is not meant to be comprehensive. The focus of this work is on summarizing the findings from the workshop and describing some interesting problems and challenges for the research community.

Note: Given the nature of the topic, we welcome and value comments and feedback on our work from the broader community. We will address the feedback in future versions of this work. Please send your comments and feedback to Mihai Christodorescu (christodorescu@google.com), Somesh Jha (jha@cs.wisc.edu), or Khawaja Shams (kshams@google.com).

References

[1] ***, *Your AI model might be telling you this is not a cat*, URL: https://art-demo.mybluemix.net/.

[2] ***, *Authors Guild letter seeks compensation from AI companies for using authors' writings in AI*, 2023. URL: https://chatgptiseatingtheworld.com/2023/07/19/authors-guild-letter-seeks-compensation-from-ai-companies-for-using-authors-writings-in-ai/.

[3] ***, *ChaosGPT: Empowering GPT with internet and memory to destroy humanity*, 2023. URL: https://www.youtube.com/watch?v=g7YJIpkk7KM.

[4] ***, *Securing the future of GenAI: Mitigating security risks*, 2023. URL: https://sites.google.com/view/genai-risk-workshop.

[5] M. S. Ackerman, "The intellectual challenge of CSCW: The gap between social requirements and technical feasibility," *Human-Computer Interaction*, vol. 15, no. 2, pp. 179–203, 2000. DOI: 10.1207/S15327051HCI1523_5.

[6] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection," in *Advanced Information Networking and Applications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020)*, Springer, pp. 1341–1354, 2020. DOI: 10.1007/978-3-030-44041-1_114.

[7] G. Alon and M. Kamfonas, *Detecting language model attacks with perplexity*, 2023.

[8] M. J. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation," in *Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4*, Springer, pp. 185–200, 2001. DOI: 10.1007/3-540-45496-9_14.

[9] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, *Constitutional AI: Harmlessness from AI feedback*, 2022. URL: https://arxiv.org/abs/2212.08073.

[10] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, and A. Szlam, *Real or fake? learning to discriminate machine from human generated text*, 2019. URL: https://arxiv.org/abs/1906.03351.

[11] A. Belanger, *OpenAI, Google will watermark AI-generated content to hinder deepfakes, misinfo*, 2023. URL: https://arstechnica.com/ai/2023/07/openai-google-will-watermark-ai-generated-content-to-hinder-deepfakes-misinfo/.

[12] M. Bohannon, *Lawyer used ChatGPT in court—And cited fake cases. a judge is considering sanctions*, 2023. URL: https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=1175e6e87c7f.

[13] A. M. Bran, S. Cox, A. D. White, and P. Schwaller, *ChemCrow: Augmenting large-language models with chemistry tools*, 2023. URL: https://arxiv.org/abs/2304.05376.

[14] B. Brittain, *Lawsuit says openai violated us authors' copyrights to train ai chatbot*, 2023. URL: https://www.reuters.com/legal/lawsuit-says-openai-violated-us-authors-copyrights-train-ai-chatbot-2023-06-29/.

[15] M. Buiten, *Product liability for defective AI*, 2023. URL: https://ssrn.com/abstract=4515202.

[16] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr, *Poisoning web-scale training datasets is practical*, 2023. URL: https://arxiv.org/abs/2302.10149.

[17] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, A. Awadalla, P. W. Koh, D. Ippolito, K. Lee, F. Tramer, and L. Schmidt, *Are aligned neural networks adversarially aligned?* 2023. URL: https://arxiv.org/abs/2306.15447.

[18] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," in *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=iC4UHbQ01Mp.

[19] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. X. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *USENIX Security Symposium*, 2020.

[20] M. Christ, S. Gunn, and O. Zamir, *Undetectable watermarks for language models*, Cryptology ePrint Archive, Paper 2023/763, 2023. URL: https://eprint.iacr.org/2023/763.

[21] DARPA Public Affairs, *DARPA announces research teams selected to semantic forensics program*, 2021. URL: https://www.darpa.mil/news-events/2021-03-02.

[22] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, *Unmasking deepfakes with simple features*, 2019. URL: https://arxiv.org/abs/1911.00686.

[23] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "Tweepfake: About detecting deepfake tweets," *PLOS ONE*, vol. 16, no. 5, pp. 1–16, 2021. URL: https://doi.org/10.1371/journal.pone.0251415.

[24] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, *The stable signature: Rooting watermarks in latent diffusion models*, 2023. URL: https://arxiv.org/abs/2303.15435.

[25] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International Conference on Machine Learning*, PMLR, pp. 3247–3258, 2020.

[26] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437, 2020. DOI: 10.1007/s11023-020-09539-2.

[27] M. Gahntz and C. Pershan, *How the eu can take on the challenge posed by general-purposeai systems*, 2022. URL: https://assets.mofoprod.net/network/documents/AI-Act_Mozilla-GPAI-Brief_Kx1ktuk.pdf.

[28] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark, *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*, 2022. URL: https://arxiv.org/abs/2209.07858.

[29] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Zhao, N. Lao, H. Lee, D.-C. Juan, and K. Guu, "RARR: Researching and revising what language models say, using language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, pp. 16477–16508, Jul. 2023. URL: https://aclanthology.org/2023.acl-long.910.

[30] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, 2020. DOI: 10.18653/v1/2020.findings-emnlp.301.

[31] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, M. R. Costa-jussà and E. Alfonseca, Eds., Florence, Italy: Association for Computational Linguistics, pp. 111–116, Jul. 2019. URL: https://aclanthology.org/P19-3019.

[32] D. Glukhov, I. Shumailov, Y. Gal, N. Papernot, and V. Papyan, *LLM censorship: A machine learning challenge or a computer security problem?* 2023. URL: https://arxiv.org/abs/2307.10719.

[33] Google, *Bard*, 2023. URL: https://bard.google.com/.

[34] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 666–667, 2020. DOI: 10.1109/CVPRW50498.2020.00341.

[35] J. He and M. Vechev, *Large language models for code: Security hardening and adversarial testing*, 2023. URL: https://arxiv.org/abs/2302.05319.

[36] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias, "TRUE: Re-evaluating factual consistency evaluation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, pp. 3905–3920, Jul. 2022. URL: https://aclanthology.org/2022.naacl-main.287.

[37] D. Hovy and D. Yang, "The importance of modeling social factors of language: Theory and practice," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 588–602, Online, 2021. DOI: 10.18653/v1/2021.naacl-main.49.

[38] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, pp. 1808–1822, Jul. 2020, URL: https://aclanthology.org/2020.aclmain.164.

[39] A. Jain, C. Adiole, S. Chaudhuri, T. Reps, and C. Jermaine, *Tuning models of code with compiler-generated reinforcement learning feedback*, 2023. URL: https://arxiv.org/abs/2305.18341.

[40] G. Jawahar, M. Abdul-Mageed, and L. Lakshmanan V.S., "Automatic detection of machine generated text: A critical survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 2296–2309, Dec. 2020. URL: https://aclanthology.org/2020.coling-main.208.

[41] Z. Jiang, J. Zhang, and N. Z. Gong, *Evading watermark based detection of AI-generated content*, 2023. URL: https://arxiv.org/abs/2305.03807.

[42] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks," in *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. URL: https://openreview.net/forum?id=eXwzgiXYM8.

[43] S. Katzenbeisser and F. Petitcolas, *Information Hiding*. Artech House, 2016.

[44] D. Kelley, *WormGPT – the Generative AI tool cybercriminals are using to launch business email compromise attacks*, 2023. URL: https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/.

[45] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, pp. 17 061–17 084, PMLR, 23–29 Jul 2023. URL: https://proceedings.mlr.press/v202/kirchenbauer23a.html.

[46] J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, and T. Goldstein, *On the reliability of watermarks for large language models*, 2023. URL: https://arxiv.org/abs/2306.04634.

[47] K. Krishna, Y. Song, M. Karpinska, J. F. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," in *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=WbFhFvjjKj.

[48] R. Krishnan, *FraudGPT: The villain avatar of ChatGPT*, 2023. URL: https://netenrich.com/blog/fraudgpt-the-villain-avatar-of-chatgpt.

[49] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang, *Robust distortion-free watermarks for language models*, 2023. URL: https://arxiv.org/abs/2307.15593.

[50] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, "GPT detectors are biased against non-native English writers," *Patterns*, vol. 4, no. 7, p. 100 779, 2023. URL: https://www.sciencedirect.com/science/article/pii/S2666389923001307.

[51] Q. V. Liao and Z. Xiao, *Rethinking model evaluation as narrowing the socio-technical gap*, 2023. URL: https://arxiv.org/abs/2306.03100.

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A robustly optimized BERT pretraining approach*, 2019. URL: https://arxiv.org/abs/1907.11692.

[53] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8060–8069, 2020.

[54] N. Lukas and F. Kerschbaum, "PTW: Pivotal tuning watermarking for pre-trained image generators," *32nd USENIX Security Symposium*, 2023.

[55] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, *Self-Refine: Iterative refinement with self-feedback*, 2023. URL: https://arxiv.org/abs/2303.17651.

[56] Makyen, *Temporary policy: Generative AI (e.g., ChatGPT) is banned*, 2022. URL: https://meta.stackoverflow.com/questions/421831/temporary-policy-generative-ai-e-g-chatgpt-is-banned.

[57] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, pp. 384–389, 2018. DOI: 10.1109/MIPR.2018.00084.

[58] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, pp. 1–6, 2019. DOI: 10.1109/WIFS47025.2019.9035099.

[59] S. McCloskey and M. Albright, *Detecting GAN-generated imagery using color cues*, 2018. URL: https://arxiv.org/abs/1812.08247.

[60] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, and N. McAleese, *Teaching language models to support answers with verified quotes*, 2022. URL: https://arxiv.org/abs/2203.11147.

[61] Meta, *Introducing Llama 2*, 2023. URL: https://ai.meta.com/llama/.

[62] Midjourney, *Midjourney*, 2023. URL: https://www.midjourney.com/.

[63] Midjourney, *Stable diffusion*, 2023. URL: https://stablediffusionweb.com/.

[64] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23, Honolulu, Hawaii, USA: JMLR.org, 2023. URL: https://arxiv.org/abs/2301.11305.

[65] L. Nataraj, T. M. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J. H. Bappy, and A. K. Roy-Chowdhury, "Detecting GAN generated fake images using co-occurrence matrices," *Electronic Imaging*, vol. 2019, no. 5, pp. 532–1, 2019. DOI: 10.2352/ISSN.2470-1173.2019.5.MWSF-532.

[66] E. Nijkamp, H. Hayashi, T. Xie, C. Xia, B. Pang, R. Meng, W. Kryscinski, L. Tu, M. Bhat, S. Yavuz, C. Xing, J. Vig, L. Murakhovska, C.-S. Wu, Y. Zhou, S. R. Joty, C. Xiong, and S. Savarese, *Long sequence modeling with XGen: A 7B LLM trained on 8K input sequence length*, 2023. URL: https://blog.salesforceairesearch.com/xgen/.

[67] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24 480–24 489, 2023. DOI: 10.1109/CVPR52729.2023.02345.

[68] OpenAI, *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses*, 2023. URL: https://openai.com/gpt-4.

[69] OpenAI, *ChatGPT: Optimizing language models for dialogue*, 2022. URL: https://openai.com/blog/chatgpt/.

[70] OpenAI, *DALL-E: Creating images from text*, 2023. URL: https://openai.com/research/dall-e.

[71] OpenAI, *GPT-4 technical report*, 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf.

[72] OpenAI, *GPT-2: 1.5b release*, 2019. URL: https://openai.com/research/gpt-2-1-5b-release.

[73] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., vol. 35, pp. 27730–27744, 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[74] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri, "Asleep at the keyboard? assessing the security of GitHub Copilot's code contributions," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 754–768, 2022. URL: https://arxiv.org/abs/2108.09293.

[75] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Conference on Empirical Methods in Natural Language Processing*, 2022. DOI: 10.18653/v1/2022.emnlp-main.225.

[76] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, *In-context retrieval-augmented language models*, 2023. URL: https://arxiv.org/abs/2302.00083.

[77] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter, "Measuring attribution in natural language generation models," *Computational Linguistics*, pp. 1–64, Aug. 2023. URL: https://doi.org/10.1162/coli_a_00486.

[78] J. Ricker, S. Damm, T. Holz, and A. Fischer, *Towards the detection of diffusion model deepfakes*, 2023. URL: https://arxiv.org/abs/2210.14571.

[79] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, *Can AI-generated text be reliably detected?* 2023. URL: https://arxiv.org/abs/2303.11156.

[80] M. Sellman, *My AI: Snapchat chatbot coaches 'girl, 13' on losing virginity*, 2023. URL: https://www.thetimes.co.uk/article/my-ai-snapchat-chatbot-coaches-girl-13-on-losing-virginity-dj7p6268b.

[81] Z. Sha, Z. Li, N. Yu, and Y. Zhang, *DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models*, 2023. URL: https://arxiv.org/abs/2210.06998.

[82] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, *The curse of recursion: Training on generated data makes models forget*, 2023. URL: https://arxiv.org/abs/2305.17493.

[83] R. G.-B. A. J. G. Sison, M. T. Daza, and E. C. Garrido-Merchán, "Chatgpt: More than a 'weapon of mass deception' ethical challenges and responses from the human-centered artificial intelligence (hcai) perspective," *International Journal of Human–Computer Interaction*, pp. 1–20, 2023. DOI: 10.1080/10447318.2023.2225931.

[84] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, *et al.*, *Release strategies and the social impacts of language models*, 2019. URL: https://arxiv.org/abs/1908.09203.

[85] I. Solaiman and C. Dennison, "Process for adapting language models to society (PALMS) with values-targeted datasets," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., vol. 34, pp. 5861–5873, 2021. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/2e855f9489df0712b4bd8ea9e2848c5a-Paper.pdf.

[86] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Curran Associates, Inc., vol. 33, pp. 3008–3021, 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[87] R. Taori and T. Hashimoto, "Data feedback loops: Model-driven amplification of dataset biases," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, pp. 33 883–33 920, PMLR, 2023. URL: https://proceedings.mlr.press/v202/taori23a.html.

[88] The White House, *Executive order on the safe, secure, and trustworthy development and use of artificial intelligence*, 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[89] The White House, *FACT SHEET: Biden–Harris administration announces national cyber workforce and education strategy, unleashing America's cyber talent*, 2023. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/31/fact-sheet-biden-harris-administration-announces-national-cyber-workforce-and-education-strategy-unleashing-americas-cyber-talent/.

[90] C. Troncoso and B. Preneel, *Detecting child sexual abuse material shouldn't be done at any cost*, 2023. URL: https://www.euronews.com/2023/07/04/detecting-child-sexual-abuse-material-shouldnt-be-done-at-any-cost.

[91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., vol. 30, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/-Paper.pdf.

[92] J. Wang, X. HU, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, W. Ye, H. Huang, X. Geng, B. Jiao, Y. Zhang, and X. Xie, "On the robustness of ChatGPT: An adversarial and out-of-distribution perspective," in *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. URL: https://openreview.net/forum?id=uw6HSkgoM29.

[93] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot . . . for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020. DOI: 10.1109/CVPR42600.2020.00872.

[94] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022. URL: https://openreview.net/forum?id=yzkSU5zdwD.

[95] M. Weiss, "Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions," *Technology Science*, vol. 2019121801, 2019.

[96] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang, "Challenges in detoxifying language models," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Eds., pp. 2447–2469, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021. URL: https://aclanthology.org/2021.findings-emnlp.210.

[97] Wikipedia contributors, *Dual-use technology—Wikipedia, the free encyclopedia*, 2023. URL: https://en.wikipedia.org/w/index.php?title=Dual-use_technology&oldid=1167047934.

[98] Wikipedia contributors, *Large language model—Wikipedia, the free encyclopedia*, 2023. URL: https://en.wikipedia.org/w/index.php?title=Large_language_model&oldid=1184758575.

[99] A. Wilson, P. Blunsom, and A. D. Ker, "Linguistic steganography on Twitter: Hierarchical language modeling with manual interaction," in *Media Watermarking, Security, and Forensics 2014*, A. M. Alattar, N. D. Memon, and C. D. Heitzenrater, Eds., International Society for Optics and Photonics, vol. 9028, pp. 9–25, SPIE, 2014. DOI: 10.1117/12.2039213.

[100] C. Xiang, *"He would still be here": Man dies by suicide after talking with AI chatbot, widow says*, 2023. URL: https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

[101] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan, *Recipes for safety in open-domain chatbots*, 2021. URL: https://arxiv.org/abs/2010.07079.

[102] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *The Eleventh International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[103] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[104] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, pp. 1–6, 2019. DOI: 10.1109/WIFS47025.2019.9035107.

[105] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds., Copenhagen, Denmark: Association for Computational Linguistics, pp. 2979–2989, Sep. 2017. URL: https://aclanthology.org/D17-1323.

[106] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang, *Provable robust watermarking for AI-generated text*, 2023. URL: https://arxiv.org/abs/2306.17439.

[107] X. Zhao, Y.-X. Wang, and L. Li, "Protecting language generation models via invisible watermarking," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23, Honolulu, Hawaii, USA: JMLR.org, 2023. URL: https://dl.acm.org/doi/10.5555/3618408.3620182.

[108] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, N. Z. Gong, Y. Zhang, and X. Xie, *PromptBench: Towards evaluating the robustness of large language models on adversarial prompts*, 2023. URL: https://arxiv.org/abs/2306.04528.

[109] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, *Can large language models transform computational social science?* 2023. URL: https://arxiv.org/abs/2305.03514.

[110] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, *Universal and transferable adversarial attacks on aligned language models*, 2023. URL: https://arxiv.org/abs/2307.15043.