Adversarial Threats in Climate AI: Navigating Challenges and Crafting Resilience

Sally Calengor, Sai Prathyush Katragadda, Joshua Steier
RAND Corporation
scalengor@rand.org[1], skatraga@rand.org, jsteier@rand.org

Abstract

The convergence of Artificial Intelligence (AI) with climate science is a double-edged sword. AI-enhanced modeling has transformative potential for the field, but it comes with new vulnerabilities, especially from adversarial machine learning. Such adversarial tactics can distort AI-driven climate models, producing misleading projections on phenomena like sea-level changes and temperature predictions. Beyond just modeling, AI-enhanced systems in resource management, conservation, and agriculture are at risk. Tampering with climate datasets can undermine decades of global research and erode trust, while adversarial misinformation campaigns can skew public discourse. Ethically, distorted data risks magnifying socio-environmental disparities. Addressing these challenges necessitates robust modeling using advanced techniques, data defense with cryptographic solutions, AI-driven infrastructure safeguards, and AI algorithms to detect and counter misinformation. Simply put, securing AI in climate science is not just a technical challenge, but a global imperative.

Introduction

As the sun sets on one of the hottest years on record, the global community must grapple with an undeniable reality: climate change isn't a distant threat, it's our present. Sweeping wildfires, unprecedented storms, and melting ice caps foreshadow this growing threat. However, new innovations, like climate artificial intelligence (AI), provide hope. This field promises not only to illuminate the path forward but also to revolutionize our approach to this existential challenge.

For decades, scientists have utilized computational models to understand and predict climatic patterns. However, the sheer complexity of our planet's climate systems has often outpaced traditional computational methods. The capabilities of AI, especially machine learning, allow us to sift through vast datasets, discern patterns and predict future climatic events. From predicting the rate of glacial melts in the Artic to optimizing renewable energy consumption, the fusion of AI with climate science is rapidly becoming the linchpin of modern environmental strategies.

The transformative potential of this combination is undeniable. AI-driven climate models can provide policymakers with the data and forecasts they need to enact timely and effective strategies to combat climate change. But AI is a double-edged sword. Its integration into climate response initiatives also introduces a host of potential vulnerabilities. Adversarial threats, where AI systems are manipulated for malicious ends, threaten to undermine the very solutions we are relying on. In this pivotal moment, the balance between the innovation's promise and its pitfalls is critical. Thus, the purpose of this paper is threefold: first, we call attention to the threat of adversarial AI in the field of climate science; second, we examine the far-reaching and potentially devastating political and practical consequences of this threat; and third, we suggest a multifaceted blueprint to build resilience and robust countermeasures.

What is Climate AI?

Climate AI, defined here as various machine learning (ML) algorithms that help predict, mitigate, and respond to climate-related challenges, is one of the latest innovations in the fight against climate change. In this field, ML algorithms prove especially useful when it comes to modeling – or the creation and use of large mathematical models run as computer simulations (Monteleoni, Schmidt, and McQuade 2013). Results from these models are critical for researchers looking to understand and forecast the Earth's climate (Monteleoni, Schmidt, and McQuade 2013). However, current analytical tools have been unable to keep pace with the rapidly growing amounts of climate data created by

[1] Corresponding author.

satellites, environmental sensors, and simulations (Monteleoni, Schmidt, and McQuade 2013). This is where ML, which uses algorithms to detect patterns in data, can advance the field.

The strength of ML algorithms lies in their ability to process enormous amounts of data quickly and efficiently (Cowls et al. 2021). Using simple methods for regression and more complicated deep learning models, climate AI incorporates data that could not be included in an automated analysis prior to modern ML. Images are one such data source that allow for more comprehensive analysis (Kim et al. 2019). In the process, these algorithms facilitate the comprehension of high-dimensional climate datasets and enable improved trend prediction and forecasting (Cowls et al. 2021). ML algorithms can also cut the costs of predictions when compared to current models, making them not only an effective alternative but a feasible one as well (Rolnick et al. 2019). The capabilities that ML models and frameworks provide for climate change will be invaluable in efforts to utilize data sources, determine climate trends, and make decisions that will positively benefit the environment moving forward.

Climate AI's Transformative Potential: Current Practice and Literature

The benefits of ML algorithms described above are already coming to life in practice, with novel models providing new techniques for evaluating systems like cloud formation, ice sheet flow, and permafrost (Rolnick et al. 2019). For example, new satellite campaigns have generated terabytes of data on the Artic and Antarctic that can provide context to global sea level rise (Rolnick et al. 2019). Researchers can now use ML models to examine this data, studying snow and sea ice properties or anticipating short-term changes in the sea ice extent (Rolnick et al. 2019). AI tools can also assist with predicting more common weather events, like heavy rain damage or wildfires (Cowls et al. 2021). Even private companies are stepping into the field with projects like AI for Earth initiative, Microsoft's five-year, $50 million project designed to support organizations and researchers using AI to tackle aspects of the climate crisis (Cowls et al. 2021). In conjunction with practical applications for the fight against climate change, AI and ML algorithms can also influence policy. These models can be used to predict outputs, like carbon emissions, based on present trends, informing the development of domestic and international policy (Mardani et. al 2020). As climate-related challenges grow, the nexus between AI and climate science will only become more important to mitigation and prevention efforts.

The benefits of climate AI and ML algorithms in climate science are manyfold. From predicting climate trends and extreme weather events to informing policy, AI can fundamentally shift how we respond to the climate crisis and provide us with the knowledge to build resilience on a national—if not global—scale.

Anticipating Shortfalls: Climate AI as a New Attack Surface

While the benefits of climate AI are vast, it is not a flawless solution. As a relatively new discipline, research and practice are only just beginning to explore potential pitfalls. For example, recent scholarship has highlighted and begun to address the levels of greenhouse gas emissions created by these systems. Large AI systems also consume copious amounts of water. For example, GPT-3 in Microsoft's U.S. data centers can use 700,000 liters of water a day for cooling (Li et al. 2023). Coupled with a complex global regulation environment in which key players, like the U.S., U.K., E.U. and China, lack consensus, these hurdles inhibit the transformative potential of climate AI.

That said, there is still one domain that research and practice have yet to explore: adversarial AI. As with any technological innovation, the benefits coincide with vulnerabilities. Because climate ML models, and ML models more generally, are focused on being able to make accurate predictions efficiently, the security of these models often becomes a secondary concern (Papernot et al. 2016). Widening the scope beyond modeling, the broader infrastructural ecosystem—including AI-driven systems like intelligent energy management, data-enhanced agricultural practices, and real-time water resource management—stands equally susceptible to adversarial threats. A well-executed adversarial attack within these interconnected systems could compromise their real-time data, leading to significant operational disruptions. The gravity of these risks cannot be understated, highlighting the importance of examining adversarial AI both in modeling and in practice.

Adversarial AI: An Overview

As explained above, the introduction of AI and ML into climate science creates a new attack surface through which adversaries can disrupt systems and critical functions (Oseni et al. 2021). Adversarial AI attacks are a sophisticated domain in which AI models are subtly and strategically manipulated to produce malicious or unintended outcomes (Goodfellow et al. 2014). Simply put, an adversary can influence the system by manipulating the model, the data, or both (Churilla, Vanhoudnos, and Beveridge 2023). In the process, the AI will either learn incorrectly, perform unexpectedly, or reveal privileged information about the model or its data (Churilla, Vanhoudnos, and Beveridge 2023).

Learning: Attacks that cause the ML algorithm to learn incorrectly target training data and foundational models.

Poisoning, a technique where triggers are input into the training data, is one example of this attack. By disrupting the data that the AI is learning from, the algorithm will adopt traits that the adversary can exploit later (Churilla, Vanhoudnos, and Beveridge 2023).

Performance*:* Attacks on the model's performance cause the ML model to behave in an unexpected way (Churilla, Vanhoudnos, and Beveridge 2023). These attacks typically occur during the training phase and result in the misidentification of patterns and skewed projections (Oseni et al. 2021).

Privacy: Attacks that cause the AI to reveal privileged information about the model fall into two categories. First, model extraction attacks give the adversary the tools they need to create a duplicate of a model that was intended to be private (Churilla, Vanhoudnos, and Beveridge 2023). Second, model inversion attacks reveal information about the data used to train the model (Churilla, Vanhoudnos, and Beveridge 2023). In both cases, the adversary gains knowledge that would otherwise remain private, giving them the tools to undermine the ML's performance.

Regardless of the type of attack, adversarial AI has the potential to seriously disrupt operational systems and generate far-reaching consequences. Already, adversarial attacks have been seen across domains, from health to transportation. Some, such as the misrecognition of a 3d printed turtle as a rifle may seem relatively mundane but even low-cost examples highlight the possibility of tricking computer vision systems (Athalye et al. 2017). More nefarious applications include using specific clothing patterns to deceive facial recognition software and the misdirection of chatbots so that the inference they conduct is offensive (Mothes 2023; Davis 2016). If left unchecked, the consequences of these adversarial attacks could range from generating misinformation to undermining the justice system.

The Threat of Adversarial AI in Climate Science

Much like the systems highlighted above, machine learning models used in the climate domain are vulnerable to adversarial attacks. The roots of this vulnerability can be traced back to the inherent structure of machine learning models, as described by Szegedy et al. in their 2014 paper "Intriguing Properties of Neural Networks."

Climate models are multifaceted, simulating the interplay of various components like the atmosphere, oceans, land, ice, and living organisms. Given the depth and breadth of such data, a machine learning model can easily be misled if its input is manipulated, even subtly. Furthermore, the data landscape in climate science poses challenges. A significant portion of this data is historical, and only a limited subset represents extreme or rare climatic events. When machine learning algorithms are trained on these constrained datasets, they are vulnerable to adversarial attacks that introduce unfamiliar patterns. Such unfamiliarity can lead the model to make erroneous predictions or interpretations.

Furthermore, the stakes in climate science are undeniably high. AI-driven insights often inform decisions spanning political, economic, and practical applications. Any deviation or misinformation can have far-reaching consequences, from misdirected policies and resource management to the decay of public trust. Potential feedback loops in the climate domain further complicate matters. An AI-driven rainfall prediction, for example, might influence water resource management, which in turn can alter the subsequent data the model ingests regarding water availability.

The "black-box" nature of many advanced AI models adds another layer of vulnerability. The decision-making processes within these models remain largely opaque, making it challenging to trace or understand their reasoning, especially when they err. The lack of transparency in these models has driven the emergent field of explainable-AI, which seeks techniques to probe and interpret the black box. The interdisciplinary facet of climate science, encompassing fields like oceanography, geology, and biology, presents unique security challenges. As knowledge is amalgamated from diverse disciplines, there may be gaps or oversights in standard security practices, especially at the confluence of these fields.

Lastly, the long-term projections inherent to climate models mean that the effects of adversarial attacks will persist for extended periods. The resulting skewed projections can thus misguide policies for years or even decades. Coupled with the global implications of climate change, adversarial attacks in this field can have international ramifications, affecting entire countries and populations.

Recent scholarship has begun to explore the effects of adversarial attacks on system operations. For example, Chen, Tan, and Zhang (2019) examine the vulnerability load forecasting to adversarial AI. Contributions like these are a critical first step in tackling the threat of adversarial AI in climate science; however, they fall short in addressing the far-reaching consequences of these attacks. Rather than merely a system malfunction or operational error, adversarial attacks on climate models and ML algorithms will have impacts on a variety of political and practical environments. The depth and breadth of these consequences are explored in the following section.

The Consequences of Adversarial AI in the Climate Space

The vulnerability of AI in climate modeling comes with high risks of skewed data and biased projections. AI that has suffered an adversarial attack may misrepresent climate

information and data, resulting in consequences ranging from ineffective weather predictions to underestimations of carbon emissions (University of Cambridge 2023). As the use of climate AI grows, the field must consider the butterfly effect of adversarial attacks and put greater emphasis on the security of these tools. Current research focuses mainly on the effects of adversarial attacks on the operational systems themselves while largely ignoring the broader consequences. However, giving the increasing demand for AI to assist in response and planning, more attention should be given to the political and practical implications of these variables.

Political Consequences

The consequences of adversarial attacks on climate AI can disrupt political stability through three pathways: misinformation and disinformation, misinformed policy, and inequity. First, adversarial attacks can fuel misinformation and disinformation campaigns. ML's ability to create idiomatic narratives, utilize multimedia resources, and write code "may decrease the risks and increase the rewards of running influence operations," (Kuper 2023). Further, these tools can increase the quality, quantity, and targeting capabilities of these campaigns (Kuper 2023). Any biased outputs would then call into question the integrity of the information space and the sanctity of climate datasets, simultaneously undermining the authority of the institution(s) that produced it. In the process, adversarial AI would contribute to issues of polarization and radicalization through the creation and spread of false or malicious information.

Second, skewed data can contribute to misinformed policy frameworks. As national and global leaders develop policies to fight climate change, they will lean on the conclusions drawn by the latest models. However, if those models have been subjected to an adversarial attack, the effectiveness and equity of these policies would be fundamentally altered. This opens the door for peer adversary nations and private entities to influence climate change policies, directing the flow of resources and accountability where they see fit.

Third, adversarial AI can exacerbate preexisting inequities. These attacks can reframe behaviors and construct narratives through the manipulation of models and data, specifically threatening vulnerable communities and regions. These areas are already susceptible to climate events also have the fewest resources, meaning that they will be the first to feel the effects of misguided policies or data misrepresentation (University of Cambridge 2023). Additionally, adversarial attacks can facilitate abuse of power on a global scale. With treaties like the Kyoto Protocol and Paris Agreement calling for greenhouse gas reductions and more climate-conscious governance, nations have incentives to manipulate data to protect their reputations and avoid accountability. The results facilitate political manipulation and, in the process, put vulnerable communities at greater risk with little to no recourse.

Practical Consequences

Alongside the political consequences described above, adversarial AI has devastating potential in the practical space. Inefficient resource allocation, flawed resilience and conservation techniques, and the disruption of emergency management response are all critical functions that could be undermined by adversarial attacks. First, AI can streamline the provisioning of resources. Xiang et al. 2021 provide one example, exploring how environmental planning for water resource management can improve water efficiency. However, if these models fall victim to an adversarial attack, resources could be sent to the wrong location or in the wrong number. In some cases, this misallocation would leave communities vulnerable to much higher death rates and recovery costs.

Second, climate AI has a key role to play in climate mitigation strategies and conservation efforts. For example. Zhou et al. 2010 created a model to measure the climate resilience of different localities. However, the misrepresentation of data in these models can lead to a failure to act, a loss of critical biodiversity, and a decay of critical infrastructure.

Lastly, adversarial attacks on climate AI tools could detrimentally affect national and international disaster response efforts. The use of AI to bolster warning systems could provide additional safeguards to ensure swift and effective responses to wildfires, hurricanes, and floods. In their paper, Fuli et al. 2016 integrate geographical information with Twitter technology to support decision-making and tsunami evacuation planning. However, if subjected to an adversarial attack, these warning systems could be delayed or subverted. And, in serious cases, slower response times can lead to devastating consequences.

Navigating the Threat Landscape: Countermeasures to Build Resilience

Given the damage that adversarial attacks can wreak on climate AI systems, it's critical the researchers and practitioners consider avenues for strengthening this technology. Several methods for combatting adversarial attacks on ML systems are discussed in the literature of various domains. The methods that may prove most effective in the climate space are discussed below.

Defensive Strategies

Adversarial training is an extremely popular approach to defending against adversarial attacks and involves injecting adversarial samples into the training data and training the

model on this dataset (Szegedy et al. 2013). Several applications have proven the success of adversarial training (Goodfellow, Shlens, Szegedy 2014, Huang et al. 2105, Tramèr et al. 2017). However, the problem lies in the inability to introduce all possible attacks into the dataset. Adversarial training has been applied in several fields of machine learning including computer vision (Goodfellow, Shlens, Szegedy 2014), cybersecurity (Kitada, Iyatomi 2021), and natural language processing (Hinton, Vinyals 2015). Several methods have been proposed to generate these adversarial samples (Szegedy et al. 2013, Goodfellow, Shlens, Szegedy 2014, Papernot et al. 2016, Kurakin, Goodfellow, Bengio 2018, Zhang et al. 2020), and were initially proposed for the computer vision domain but have been extended to other fields such as question answering (Tripathi, Mishra 2022) and environmental sound classification (Abou Khamis, Shafiq, Matrawy 2020). There are several other methods discussed in the literature that directly involve the data such as Gradient Hiding (Tramèr et al. 2017), Blocking the Transferability (Hosseini et al. 2017), Data Compression (Dziugaite et al. 2016), and Data Randomization (Xie et al. 2017), but these are less prominent methods and as such we do not delve into their specifics.

Distillation is a process in which learned parameters from a larger model are used to train a smaller model (Papernot et al. 2016). Defensive Distillation, proposed in Soll et al. 2019, is where the learned parameters of a deep neural network (DNN) are used to increase it robustness with respect to adversarial attacks. The authors show that on the MNIST dataset the success rate of adversarial examples was reduced from 95.89% to 0.45% and on the CIFAR-10 dataset from 87.89% to 5.11% (Soll et al. 2019). This method has successfully been extended to several domains including natural language processing (Apruzzese et al. 2020) and cybersecurity (Sagi and Lior 2018).

Ensemble methods involve using multiple machine learning models or statistical techniques for some task (Apruzzese et al. 2020). The authors of Joshi et al. 2021 propose an ensemble method named AppCon which was able to prevent 75% of adversarial attacks in their setting of data from social media applications. Ensemble methods have also been successfully employed to mitigate the threat of adversarial examples within audio data (Pawlicki, Choraś, Kozik, 2020).

Several methods exist for detecting adversarial samples prior to them ever reaching the model. Supervised and unsupervised learning strategies can be used to detect an adversarial sample prior to its contact with the machine learning model, such as in Van Tuinen et al. 2022 where k Nearest Neighbors and Random Forest are able to achieve a 99% recall in detecting adversarial machine learning samples. Distance based methods can be used to detect outliers, such as in Grosse et al. 2017, and these samples can be assumed to be adversarial examples. Lastly, a statistical technique is proposed by Treen, Williams, and O'Neill 2020 to detect adversarial examples using Maximum Mean Discrepancy and Energy Distance.

Misinformation

Misinformation related to climate change is already rampant in public discourse (Treen, Hywel, O'Neill. 2020). While we touched on the role of AI in generating and disseminating misinformation earlier in the paper, it is equally as important to examine how misinformation could affect climate AI systems. For example, natural language processing systems used in climate machine learning could be affected by statements and papers denying climate change trends. Methods of detection such as those proposed in Van Tuinen et al. 2022, Grosse et al. 2017, and Treen, Hywel, and O'Neill 2020, along with the many methods presented in Dhiman et al. 2023 may prove beneficial in ensuring that this misinformation does not reach the model.

Possible Applications in Climate AI

Using adversarial training methods will prove critical in developing and validating robust climate AI models and algorithms. These efforts should be coupled with detection methods to ensure that the model is not fed adversarial data during the deployment life cycle.

To better illustrate how the mentioned methods might benefit climate ML, we provide results from a simple experiment. We use the climate change dataset from Kaggle to predict atmospheric concentration of CO2 given several other variables. To do so, a random forest regression model is used from scikit-learn (Pedregosa et al. 2011), and we split the dataset such that 20% is reserved for testing. The initial performance of the model reaches roughly 97% accuracy in predicting concentration of CO^2. To simulate an adversarial attack, we add random noise drawn from a gaussian distribution with mean 0 and varying levels of sigma to each of the features in 10% of the testing dataset. We then

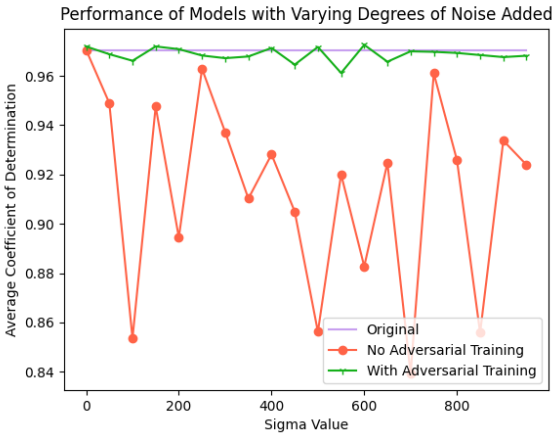

Figure 1: Effect of Adversarial Training on Model Performance

run the previously trained model on this corrupted dataset to determine performance. Then, to simulate adversarial training, we conduct the same perturbation on 10% of the training data, train a random forest regressor on this dataset, and record the performance on the corrupted testing set. We iterate this experiment over varying levels of sigma, ranging from 0 to 1000 in increments of 50. A plot comparing performances is shown in Figure 1, where 'original' refers to the model trained and tested on the uncorrupted data. This original data is included to show the natural variance associate with a random forest model. As shown in the figure, the model with no adversarial training was deeply affected by the introduction of random noise to each of the features. The model with adversarial training, on the other hand, was much more resilient to the attack. Though this experiment is basic, it does show the extreme effect that adversarial training can have in a climate ML application.

Conclusion

From predicting the rate of global temperature changes to assisting with evacuation planning, the use of climate AI has become a critical tool in the movement to mitigate climate change. AI-driven climate models provide unparalleled insights into environmental trends, and they equip policymakers with the essential data and forecasts required to devise effective strategies. Yet, as with all tools, climate AI is not infallible.

As we've explored in this paper, the integration of these models into our climate response initiatives presents new vulnerabilities. The complexity of the systems, coupled with the inherent intricacies of machine learning models, makes them susceptible to adversarial attacks. Such attacks aim to manipulate AI systems for malicious purposes, posing significant threats to the models we increasingly rely on. The repercussions of successful adversarial attacks extend beyond mere system malfunctions. They can destabilize political landscapes, erode public trust, and critically impair our collective ability to respond to climate emergencies. The prospect of such dire consequences underscores the imperative to safeguard our AI tools against these threats.

Our position is clear: the research community must intensify its focus on investigating adversarial AI vulnerabilities specific to climate models. Through a combination of proactive "red teaming" exercises – where experts simulate adversarial attacks to uncover weaknesses – and rigorous recommendations, we can anticipate potential threats and devise strategies to counteract them.

The benefits of climate AI are too great to be undermined by vulnerabilities. By dedicating resources to understanding, identifying, and mitigating potential threats, we can ensure that our AI tools remain both effective and resilient. Future research endeavors should prioritize the robust countermeasures discussed herein, striving for a future where climate AI remains reliable, accessible, and, above all, trustworthy.

References

Abou Khamis R.; Shafiq, M.O.; Matrawy, A. 2020. Investigating Resistance of Deep Learning-Based IDS Against Adversaries Using Min-Max Optimization. arXiv preprint. arXiv: arXiv:1910.14107 (cs.LG). https://doi.org/10.48550/arXiv.1910.14107.

Ai, F.; Comfort, K.; Dong, Y.; Znati, T. 2016. A Dynamic Decision Support System Based on Geographical Information and Mobile Social Networks: A Model for Tsunami Risk Mitigation in Padang, Indonesia. Safety Science 90: 62-74. https://www.sciencedirect.com/science/article/abs/pii/S0925753515002519.

Apruzzese, G.; Andreolini, M.; Colajanni, M.; Marchetti, M. 2019. Hardening Random Forest Cyber Detectors Against Adversarial Attacks. arXiv preprint. arXiv:1912.03790 (cs.CR). https://doi.org/10.48550/arXiv.1912.03790.

Apruzzese, G.; Andreolini, M.; Marchetti, M.; Colacino, V.G.; Russo, G. AppCon: Mitigating Evasion Attacks to ML Cyber Detectors. Symmetry 12(4): 653. https://doi.org/10.3390/sym12040653.

Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. 2017. Synthesizing Robust Adversarial Examples. arXiv preprint. arXiv:1707.07397 (cs.CV).

Chen, Y.; Tan, Y.; Zhang, B. 2019. Exploiting Vulnerabilities of Load Forecasting Through Adversarial Attacks. arXiv preprint. arXiv:1904.06606 (cs. SY).

Churilla, M.; Vanhoudnos, N.; Beveridge, R. May 15, 2023. "The Challenge of Adversarial Machine Learning." SEI Blog. As of September 10, 2023: https://insights.sei.cmu.edu/blog/the-challenge-of-adversarial-machine-learning/.

Cowls, J.; Tsamados, A.; Teddeo, M.; Floridi, L. 2021. The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change—Opportunities, Challenges, and Recommendations. AI & Society 38: 283-307. 10.1007/s00146-021-01294-x.

Davis, E. 2016. AI Amusements: The Tragic Tale of Tay the Chatbot. AI Matters 2(4): 20-24. http://dx.doi.org/10.1145/3008665.3008674.

Dhiman, P.; Kaur, A.; Iwendi, C.; Kumar Mohan, S. A Scientometric Analysis of Deep Learning Approaches for Detecting Fake News. Electronics 12(4): 948. https://doi.org/10.3390/electronics12040948.

Dziugaite, G.; Ghahramani, Z.; Roy, D. 2016. A Study of the Effect of JPG Compression on Adversarial Images. arXiv preprint. arXiv:1608.00853 (cs.CV). https://doi.org/10.48550/arXiv.1608.00853.

Goodfellow, I.; Shlens, J.; Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. arXiv preprint. arXiv:1412.6572 (stat.ML).

Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; McDaniel, P. 2017. On The (Statistical) Detection of

Adversarial Examples. arXiv preprint. arXiv: 1702.06280 (cs.CR). https://doi.org/10.48550/arXiv.1702.06280.

Hinton, G.; Vinyals, O.; Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv preprint. arXiv:1503.02531 (stat.ML). https://doi.org/10.48550/arXiv.1503.02531.

Hosseini, H.; Chen, Y.; Kannan, S.; Zhang, B.; Poovendran, R. 2017. Blocking Transferability of Adversarial Examples in Black-Box Learning Systems. arXiv preprint. arXiv:1703.04318 (cs.LG). https://doi.org/10.48550/arXiv.1703.04318.

Huang, R.; Xu, B.; Schuurmans, D.; Szepescvari, C. 2016. Learning With a Strong Adversary. arXiv preprint. arXiv:1511.03034 (cs. LG).

Jobin, A.; Ienca, M.; Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence 1: 389-399. https://doi.org/10.1038/s42256-019-0088-2.

Joshi, S.; Villalba, J.; Żelasko, P.; Moro-Velázquez, L.; Dehak, N. 2021. Study of Pre-Processing Defenses Against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems, arXiv preprint. arXiv:2101.08909 (eess.AS). https://doi.org/10.48550/arXiv.2101.08909.

Kim, S.; Kim, H; Lee, J.; Yoon, S.; Kahou, S.; Kashinath, K.; Prabhat, M. 2019. Deep-Hurricane-Tracker: Tracking and Forecasting Extreme Climate Events. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). doi: 10.1109/WACV.2019.00192.

Kitada, S., Iyatomi, H. 2021. Attention Meets Perturbations: Robust and Interpretable Attention with Adversarial Training. arXiv preprint. arXiv:2009.12064 (cs.CL). https://doi.org/10.48550/arXiv.2009.12064.

Kuper, A. August 23, 2023. Generative AI and Disinformation. Bipartisan Policy Center. As of September 10, 2023: https://bipartisanpolicy.org/blog/generative-ai-and-disinformation/.

Kurakin, A.; Goodfellow, I.J.; Bengio, S. 2017. Adversarial Examples in the Physical World. arXiv preprint. arXiv:1607.02533 (cs.CV). https://doi.org/10.48550/arXiv.1607.02533.

Mardani, A.; Huchang, L.; Nilashi, M.; Alrasheedi, M.; Cavallaro, F. 2020. A Multi-State Method to Predict Carbon Dioxide Emissions Using Dimensionality Reduction, Clustering, and Machine Learning Techniques. Journal of Cleaner Production 275. https://doi.org/10.1016/j.jclepro.2020.122942.

Monteleoni, C.; Schmidt, G.; McQuade, S. 2013. Climate Informatics: Accelerating Discovering in Climate Science with Machine Learning. Computing in Science and Engingeering. September/October 2013.

Mothes, Kate. “Trick Facial Recognition Software into Thinking You’re a Zebra or Giraffe with These Pyschedelic Garments. Colossal,1Feb.2023,www.thisiscolossal.com/2023/02/capable-facial-recognition textiles/#:~:text=Evocative%20of%20Magic%20Eye%20puzzles,time%2C%E2%80%9D%20the%20company%20says.

Oseni, A.; Moustafa, N.; Janicke, H.; Liu, P.; Tari, Z.; Vasilakos, A. 2021. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. arXiv preprint. arXiv:2102.04661 (cs.CR). https://doi.org/10.48550/arXiv.2102.04661.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. 2015. The Limitations of Deep Learning in Adversarial Settings. arXiv preprint. arXiv:1511.07528 (csCR). https://doi.org/10.48550/arXiv.1511.07528.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy. arXiv:1511.04508v2 (cs.CR).

Pedregosa, F.; Varoquax, G.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duschesnay, É. 2011. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 12: 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

piAI & Kaggle (Eds.). (n.d.). climate change [Dataset]. https://www.kaggle.com/datasets/econdata/climate-change?resource=download

Rolnick, D.; Donti, P. L.; Kaack, L. H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; Luccioni, A.; Maharaj, T.; Sherwin, E.; Mukkavilli, S.; Kording, K.; Gomes, C.; Ng, A.; Hassabis, D.; Platt, J.; Creutzig, F.; Chayes, J.; Bengio, Y. 2019. Tackling Climate Change with Machine Learning. arXiv preprint. arXiv:1906.05433.

Sagi, O., Rokach, L. 2018. Ensemble Learning: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4). https://doi.org/10.1002/widm.1249.

Schwartz, R.; Dodge, J.; Smith, N.; Etzioni, O. 2020. Green AI. Communications of the ACM 63(12): 54-63. 10.1145/3381831.

Soll, M.; Hinz, T.; Magg, S.; Wermter, S. 2019. Evaluating Defensive Distillation for Defending Text Processing Neural Networks Against Adversarial Examples. arXiv preprint. arXiv:1908.07899 (cs.CL). https://doi.org/10.48550/arXiv.1908.07899.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Dumitru, E.; Goodfellow, I.; Fergus, R. 2013. Intriguing Properties of Neural Networks. arXiv preprint. arXiv:1312.6199 (cs.CV).

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. 2017. Ensemble Adversarial Training: Attacks and Defenses. arXiv preprint. arXiv:1705.07204 (stat.ML).

Treen, K.; Treen, d'I.; Hywel, W.; O'Neill, S. 2020. Online Misinformation About Climate Change. Wiley Interdisciplinary Reviews: Climate Change 11(5). https://doi.org/10.1002/wcc.665.

Tripathi, A.M., Mishra, A. 2022. Adv-ESC: Adversarial Attack Datasets for an Environmental Sound Classification. Applied Acoustics 185. https://doi.org/10.1016/j.apacoust.2021.108437.

Van Tuinen, J.; Ranganath, A.; Konjevod, G.; Singhal, M.; Marcia, R. 2022. Novel Adversarial Defense Techniques for White-box Attacks. In Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2022, pp. 617–622.

Xiang, X.;, Li, Q.; Khan, S.; Ibrahim Khalaf, O. 2021. Urban Water Resource Management for Sustainable Environmental Planning Using Artificial Intelligence Techniques. Environmental Impact Assessment Review 86. https://doi.org/10.1016/j.eiar.2020.106515.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. 2017. Mitigating Adversarial Effects Through Randomization. arXiv preprint. arXiv:1711.01991 (cs.CV). https://doi.org/10.48550/arXiv.1711.01991.

Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. 2019. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. ACM Transitions on Intelligent Systems and Technology 11(3): 1–41. https://doi.org/10.1145/3374217.

Zhou, H.; Wang, J.; Wan, J.; Jia, H. 2010. Resilience to Natural Hazards: A Geographic Perspective. Natural Hazards 53: 21-41. https://link.springer.com/article/10.1007/s11069-009-9407-y.