# How Robust Are Energy-Based Models Trained With Equilibrium Propagation?

**Siddharth Mansingh, Michal Kucer, Garrett Kenyon, Juston Moore & Michael Teti**
Los Alamos National Laboratory
Los Alamos, NM 87545
{smansingh, michal, gkenyon, jmoore01, mteti}@lanl.gov

## Abstract

Deep neural networks (DNNs) are easily fooled by adversarial perturbations that are imperceptible to humans. Adversarial training, a process where adversarial examples are added to the training set, is the current state-of-the-art defense against adversarial attacks, but it lowers the model's accuracy on clean inputs, is computationally expensive, and offers less robustness to natural noise. In contrast, energy-based models (EBMs), which were designed for efficient implementation in neuromorphic hardware and physical systems, incorporate feedback connections from each layer to the previous layer, yielding a recurrent, deep-attractor architecture which we hypothesize should make them naturally robust. Our work is the first to explore the robustness of EBMs to both natural corruptions and adversarial attacks, which we do using the CIFAR-10 and CIFAR-100 datasets. We demonstrate that EBMs are more robust than transformers and display comparable robustness to adversarially-trained DNNs on white-box, black-box, and natural perturbations without sacrificing clean accuracy, and without the need for adversarial training or additional training techniques.

## 1 Robustness with Top-Down/Feedback Connections

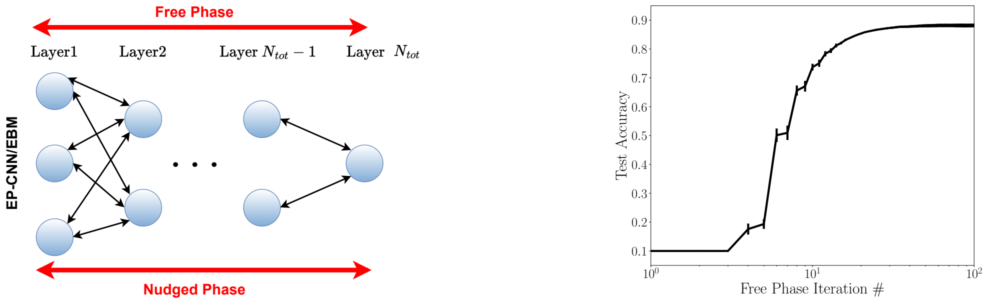
Deep neural networks (DNNs) are non-robust to manipulated inputs that are imperceptible to humansSzegedy et al. [2014], Madry et al. [2017], as well as natural noise Hendrycks and Dietterich [2019]. The current state-of-the-art defense against adversarial attacks Madry et al. [2017] involves training on adversarial examples. However, adversarial training leads to a drop in accuracy on clean/unperturbed test input Tsipras et al. [2018], a well-established tradeoff that has been described theoretically Schmidt et al. [2018], Zhang et al. [2019] and observed experimentally Stutz et al. [2019], Raghunathan et al. [2019]. Moreover, adversarially-trained models overfit to the attack they are trained with and perform poorly under different attacks Wang et al. [2020], as well as natural noise/corruptions. On the other hand, ViTs have shown increased robustness compared to standard Convolutional Neural Networks (CNNs) without requiring adversarial training. However, ViTs require very large datasets containing millions of samples or more to achieve good clean and robust accuracy Lee et al. [2022], which is simply not realistic in many applications. Training large models as well as adversarial training runs into problems of huge energy consumption that is not environmentally sustainable as well as inaccessible to the majority of researchers.

Biological perceptual systems, in contrast, are much more robust to noise and perturbations, can learn from much fewer examples, not suffer from any drop in clean accuracy, and require much less power than standard DNNs. One reason for this discrepancy in performance/behavior is the fact that DNNs lack many well-known structural motifs present in biological sensory systems. For example, feedback connections are abundant in virtually every sensory area of mammals Erişir et al. [1997], Ghazanfar et al. [2001], Boyd et al. [2012], Homma et al. [2017], Jin et al. [2021], often

Associative Memory & Hopfield Networks in 2023. NeurIPS 2023 workshop.

outnumbering feedforward connections by many times Van Essen and Maunsell [1983], Erişir et al. [1997], yet they remain absent in the vast majority of DNNs. Ample neuroscientific evidence suggests that these massive feedback networks convey rich information from higher to lower cortical areas, such as sensory context, Angelucci and Bressloff [2006], Czigler and Winkler [2010], Angelucci et al. [2017], top-down attention Luck et al. [1997], Noudoost et al. [2010], and expectation Rao and Ballard [1999]. It is also thought that feedback is critical for reliable inference from weak or noisy stimuli DiCarlo et al. [2012], especially in real-world or complex scenarios with competing stimuli Desimone and Duncan [1995], Kastner and Ungerleider [2001], McMains and Kastner [2011], Homma et al. [2017].

## 2 Robustness of Equilibrium Propagation



(a) In EP-CNNs, learning (nudged phase) as well as inference (free phase) are dynamic recurrent processes where information travels bidirectionally through feedforward and feedback connections.

(b) Test accuracy achieved by a trained **EP-CNN** model, evaluated at each iteration of the free-phase. Error bars represent five models trained with five different seeds.

Figure 1: Characteristics of the Equilibrium Propagation learning framework

We hypothesize that, when incorporated into standard DNNs, top-down feedback will lead to increased robustness against adversarial attacks and natural corruptions on standard image recognition tasks. To investigate this, we focus on a recent class of biologically-plausible DNNs referred to as Energy-Based Models (EBMs), which are trained with a learning framework referred to as Equilibrium Propagation (EP). In contrast to standard DNNs, information in EBMs flows both forward and backward due to the incorporation of feedback connections between consecutive layers. These feedback connections allow EBMs to be trained with a spatio-temporally local update rule (EP) Scellier and Bengio [2017], Luczak et al. [2022] in low-power neuromorphic hardware, an important factor given the environmental cost of training DNNs on standard hardware Strubell et al. [2020], Van Wynsberghe [2021]. This feedback also endows EBMs with global attractors, which should make EBMs more robust to perturbations. After the landmark study by Scellier and Bengio [2017], recent advances in EP have focused primarily on scaling EBMs to larger and more complex tasks Laborieux et al. [2021], Kubo et al. [2022] or modifying them for use in non-standard hardware Kendall et al. [2020], Laborieux and Zenke [2022]. As a result, studies on the robustness of EP-based models to adversarial and natural perturbations remain nonexistent.

## 3 Related Works on Adversarial Defenses

The use of recurrent networks that involve complex dynamics to reach a steady state is common in biologically plausible defense methods. Inclusion of trainable feedback connections inspired by regions in cerebral cortex Walsh et al. [2020], in order to implement predicting coding frameworks, have demonstrated marginal robustness against both black-box and white-box attacks Boutin et al. [2020], Choksi et al. [2021]. Evidence of perceptual straightening of natural movie sequences in human visual perception Hénaff et al. [2019] has also inspired robust perceptual DNNs which integrate visual information over time, leading to robust image classification Vargas et al. [2020], Daniali and Kim [2023]. However, the above models have either not been tested against adversarial or natural perturbations, and the ones that have, have only exhibited marginal increases in robustness relative to standard DNNs.

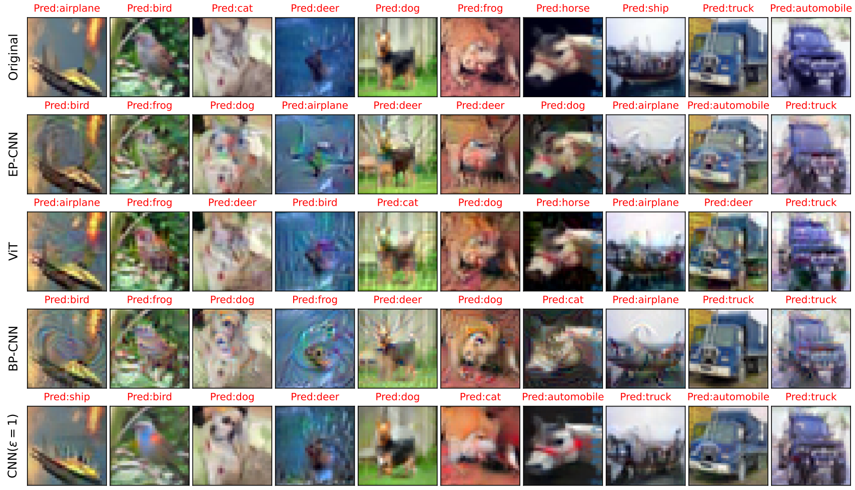
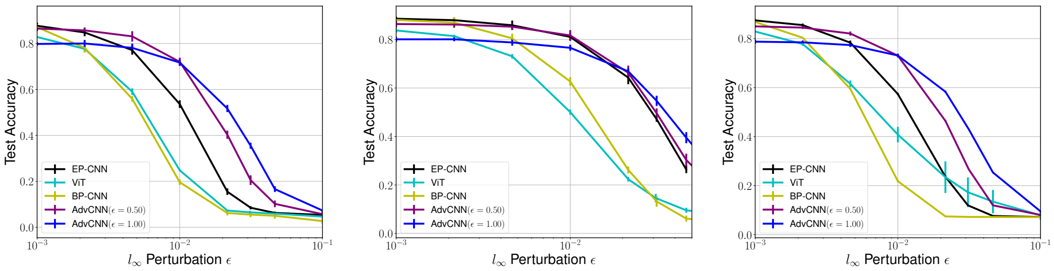


Figure 2: Examples of perturbed images using an $l_2$ PGD attack with $\epsilon = 3$. Attacks on EBMs/EP appear more semantic compared to other models.

A line of work similar to equilibrium propagation was introduced by Bai et al. [2019], known as deep equilibrium models (DEQ). DEQs involve finding fixed points of a single layer and since the fixed point can be thought of as a local attractor, these models were expected to be robust to small input perturbations, although empirical observations have proven otherwise Gurumurthy et al. [2021]. Robustness evaluations of DEQs often involve approximate/inexact gradients in order to carry out gradient-based attacks, raising concerns about gradient obfuscation Liang et al. [2021], Wei and Kolter [2021], Yang et al. [2022]. Also, DEQs alone are not robust to adversarial attacks, and, as a result, are often paired with adversarial training or other additional techniques to gain robustness Li et al. [2022], Chu et al. [2023], Yang et al. [2023].

# 4 Attacks on Equilibrium Propagation



(a) AutoAttack

(b) Square Attack

(c) PGD Attack

Figure 3: Adversarial attack results on CIFAR-10 dataset. Error bars represent the 95% CI over 5 different seeds.

For our experimental setup, we trained a model (referred to as **EP-CNN**) with equilibrium propagation using the symmetric weight update rule provided by Laborieux et al. [2021]. **EP-CNN** had four convolutional layers followed by a fully connected (FC) layer. To compare the performance of EP with standard training techniques, we also trained a model consisting of four convolutional layers followed by a fully connected layer with backpropagation, referred to as **BP-CNN**. Additionally each of the convolutional layers was followed by a batch normalization layer. While these **BP-CNN**s' did not achieve state-of-the-accuracy on the tested datasets, the purpose was to compare results from EP with an equivalent model. We also adversarially trained a model with similar architecture as that of **BP-CNN** with various $\|\epsilon_2\|$ constraints and 200 iterations of the projected gradient descent (PGD) attacks Madry et al. [2017]. Finally we also trained an vision transformer **ViT** with 7 layers, each with 12 heads using a patch size of $4 \times 4$. These architectural hyperparameters were chosen to provide a clean test performance similar to that of other models, using a grid search.

We evaluated the adversarial robustness of Equilibrium Propagation, a biologically plausible learning framework, compatible with neuromorphic hardware, for image classification tasks. We demonstrate clean accuracy comparable to models trained with backpropagation. Through our experiments, we demonstrate competitive accuracy and inherent adversarial robustness of **EP-CNN**s to natural corruptions and black-box attacks. We also demonstrate competitive robustness to white-box attacks when compared with adversarially-trained models **AdvCNN**. **EP-CNN**s far outperform the **ViT** models across the datasets used in this study for both adversarial and natural noise, even though **ViT** models were trained on input extensively augmented using similar noise perturbations. **EP-CNN**s also do not suffer from lower clean accuracy unlike models that have been adversarially trained. These adversarially trained models also fail catastrophically when subjected to noise they were not trained on, such as the natural corruptions, whereas **EP-CNN**s are far more robust across both adversarial attacks and natural corruptions, without any extensive augmentation or adversarial training.

## 5 Discussion and Conclusion

The role of feedback connections in the brain has long been overlooked in DNNs. Neuroscientific studies suggest that the abundant feedback connections present in the cortex are not merely modulatory and convey valuable information from higher to lower cortical areas, such as sensory context and top-down attention. Recent experiments have shown that time-limited humans process adversarial images much differently compared to their DNN counterparts Elsayed et al. [2018], thus leading to the hypothesis that perception of static images is a dynamic process and benefits hugely from recurrent feedback connections Daniali and Kim [2023]. Earlier studies Hupé et al. [1998] hypothesized the role of feedback connections in discriminating the object of interest from background information and recent studies Kar et al. [2019] showed that challenging images took more time to be recognized compared to control images, providing more reasons to believe that feedback connections is critical to improving robustness of the ventral stream. Our findings solidify the above claim, thus paving the way for robust artificial networks that include feedback connections.

While **EP** is not the only training algorithm that involves settling into a fixed point before making inference, the complex dynamics showcased in **EP** gets rid of small input perturbations in the process of attaining a steady state. In case of white-box attacks, large perturbations computed on EBMs/EP appear *semantically meaningful* as shown in Figure 2 in contrast to all other models tested, thus strengthening the hypothesis that large perturbations are able to move the trajectory of the state past a decision boundary for models trained with EP. This is concurrent with the assumption that equilibrium propagation allows learning of features of the input dataset in a hierarchical manner, akin to the hierarchical ground state structure of a spin-glass.

As mentioned in the initial paper Scellier and Bengio [2017], we also found that **EP** is relatively sensitive to the hyperparameters used to train the model as well as the seed used to initialize the weights. Since the inference is defined implicitly in terms of the input and the parameters of the model, this makes even our optimized implementation less practical for applications on traditional hardware (like GPUs). Apart from the instability of EP models shown during training, another limitation of our work is the amount of time required to perform the free phase to reach a steady state, when trained on traditional GPUs. While this limits the EP models to relatively shallow architectures and datasets when using standard hardware, EP-CNNs are ideal for implementation in neuromorphic hardware, leading to faster and more robust bio-plausible models. While spiking implementations of EP exist O'Connor et al. [2019], future work would then involve optimizing those implementations in order to run on realistic datasets like ImageNet.

We performed the first investigation into the robustness of EBMs trained with EP (EP-CNNs) to adversarial and natural perturbations/noise. Our results indicate that EP-CNNs are significantly more robust than standard CNNs and ViTs. We also show that EP-CNNs exhibit significantly greater robustness to natural perturbations and similar robustness to state-of-the-art black-box attacks when compared with adversarially-trained CNNs, but they do not suffer from decreased accuracy on clean data. We also find that the adversarial attacks on EBMs are more semantic than those computed on standard and adversarially-trained DNNs, which indicates that EBMs learn features that are truly useful for the classification tasks they are trained on. Overall, our work indicates that many of the problems exhibited by current DNNs, including poor energy-efficiency and robustness, can be solved by an elegant, biophysically-plausible framework for free.

# 6 Acknowledgements

SM would like to thank the Center for Nonlinear Studies and Cyber Toaster Summer School at Los Alamos National Laboratory (LANL) for supporting this work via a graduate fellowship. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy (Contract No. 89233218CNA000001).

## References

A. Angelucci and P. C. Bressloff. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. *Progress in brain research*, 154:93–120, 2006.

A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, and P. C. Bressloff. Circuits and mechanisms for surround modulation in visual cortex. *Annual review of neuroscience*, 40:425–451, 2017.

S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

V. Boutin, A. Franciosini, F. Ruffier, and L. Perrinet. Effect of top-down connections in hierarchical sparse coding. *Neural Computation*, 32(11):2279–2309, Nov. 2020. doi: 10.1162/neco_a_01325. URL `https://doi.org/10.1162/neco_a_01325`.

A. M. Boyd, J. F. Sturgill, C. Poo, and J. S. Isaacson. Cortical feedback control of olfactory bulb circuits. *Neuron*, 76(6):1161–1174, 2012.

B. Choksi, M. Mozafari, C. Biggs O'May, B. Ador, A. Alamia, and R. VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Advances in Neural Information Processing Systems*, 34:14069–14083, 2021.

H. Chu, S. Wei, and T. Liu. Learning robust deep equilibrium models. *arXiv preprint arXiv:2304.12707*, 2023.

I. Czigler and I. Winkler, editors. *Unconscious Memory Representations in Perception*. John Benjamins Publishing Company, May 2010. doi: 10.1075/aicr.78. URL `https://doi.org/10.1075/aicr.78`.

M. Daniali and E. Kim. Perception over time: Temporal dynamics for robust image understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 5656–5665. IEEE, 2023. doi: 10.1109/CVPRW59228.2023.00599. URL `https://doi.org/10.1109/CVPRW59228.2023.00599`.

R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, Mar. 1995. doi: 10.1146/annurev.ne.18.030195.001205. URL `https://doi.org/10.1146/annurev.ne.18.030195.001205`.

J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, Feb. 2012. doi: 10.1016/j.neuron.2012.01.010. URL `https://doi.org/10.1016/j.neuron.2012.01.010`.

G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

A. Erişir, S. C. Van Horn, M. E. Bickford, and S. M. Sherman. Immunocytochemistry and distribution of parabrachial terminals in the lateral geniculate nucleus of the cat: a comparison with corticogeniculate terminals. *Journal of Comparative Neurology*, 377(4):535–549, 1997.

A. A. Ghazanfar, D. J. Krupa, and M. A. Nicolelis. Role of cortical feedback in the receptive field structure and nonlinear response properties of somatosensory thalamic neurons. *Experimental brain research*, 141:88–100, 2001.

S. Gurumurthy, S. Bai, Z. Manchester, and J. Z. Kolter. Joint inference and input optimization in equilibrium networks. *Advances in Neural Information Processing Systems*, 34:16818–16832, 2021.

O. J. Hénaff, R. L. T. Goris, and E. P. Simoncelli. Perceptual straightening of natural videos. *Nature Neuroscience*, 22(6):984–991, Apr. 2019. doi: 10.1038/s41593-019-0377-4. URL `https://doi.org/10.1038/s41593-019-0377-4`.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

N. Y. Homma, M. F. Happel, F. R. Nodal, F. W. Ohl, A. J. King, and V. M. Bajo. A role for auditory corticothalamic feedback in the perception of complex sounds. *Journal of Neuroscience*, 37(25): 6149–6161, 2017.

J. M. Hupé, A. C. James, B. R. Payne, S. G. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, 394 (6695):784–787, Aug. 1998. doi: 10.1038/29537. URL `https://doi.org/10.1038/29537`.

H. Jin, Z. H. Fishman, M. Ye, L. Wang, and C. S. Zuker. Top-down control of sweet and bitter taste in the mammalian brain. *Cell*, 184(1):257–271, 2021.

K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, Apr. 2019. doi: 10.1038/s41593-019-0392-5. URL `https://doi.org/10.1038/s41593-019-0392-5`.

S. Kastner and L. G. Ungerleider. The neural basis of biased competition in human visual cortex. *Neuropsychologia*, 39(12):1263–1276, Jan. 2001. doi: 10.1016/s0028-3932(01)00116-6. URL `https://doi.org/10.1016/s0028-3932(01)00116-6`.

J. Kendall, R. Pantone, K. Manickavasagam, Y. Bengio, and B. Scellier. Training end-to-end analog neural networks with equilibrium propagation. *arXiv preprint arXiv:2006.01981*, 2020.

Y. Kubo, E. Chalmers, and A. Luczak. Combining backpropagation with equilibrium propagation to improve an actor-critic reinforcement learning framework. *Frontiers in Computational Neuroscience*, 16:980613, 2022.

A. Laborieux and F. Zenke. Holomorphic equilibrium propagation computes exact gradients through finite size oscillations. *Advances in Neural Information Processing Systems*, 35:12950–12963, 2022.

A. Laborieux, M. Ernoult, B. Scellier, Y. Bengio, J. Grollier, and D. Querlioz. Scaling equilibrium propagation to deep ConvNets by drastically reducing its gradient estimator bias. *Frontiers in Neuroscience*, 15, Feb. 2021. doi: 10.3389/fnins.2021.633674. URL `https://doi.org/10.3389/fnins.2021.633674`.

S. Lee, S. Lee, and B. C. Song. Improving vision transformers to learn small-size dataset from scratch. *IEEE Access*, 10:123212–123224, 2022. doi: 10.1109/access.2022.3224044. URL `https://doi.org/10.1109/access.2022.3224044`.

M. Li, Y. Wang, and Z. Lin. CerDEQ: Certifiable deep equilibrium model. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12998–13013. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/li22t.html`.

K. Liang, C. Anil, Y. Wu, and R. Grosse. Out-of-distribution generalization with deep equilibrium models. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of neurophysiology*, 77(1): 24–42, 1997.

A. Luczak, B. L. McNaughton, and Y. Kubo. Neurons learn by predicting future activity. *Nature machine intelligence*, 4(1):62–72, 2022.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

S. McMains and S. Kastner. Interactions of top-down and bottom-up mechanisms in human visual cortex. *The Journal of Neuroscience*, 31(2):587–597, Jan. 2011. doi: 10.1523/jneurosci.3766-10.2011. URL `https://doi.org/10.1523/jneurosci.3766-10.2011`.

B. Noudoost, M. H. Chang, N. A. Steinmetz, and T. Moore. Top-down control of visual attention. *Current Opinion in Neurobiology*, 20(2):183–190, Apr. 2010. doi: 10.1016/j.conb.2010.02.003. URL `https://doi.org/10.1016/j.conb.2010.02.003`.

P. O'Connor, E. Gavves, and M. Welling. Training a spiking neural network with equilibrium propagation. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1516–1523. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/o-connor19a.html`.

A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

B. Scellier and Y. Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11, May 2017. doi: 10.3389/fncom.2017.00024. URL `https://doi.org/10.3389/fncom.2017.00024`.

L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020.

D. Stutz, M. Hein, and B. Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL `http://arxiv.org/abs/1312.6199`.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

D. C. Van Essen and J. H. Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6:370–375, 1983.

A. Van Wynsberghe. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1 (3):213–218, 2021.

D. V. Vargas, B. Liao, and T. Kanzaki. Perceptual deep neural networks: Adversarial robustness through input recreation. *arXiv preprint arXiv:2009.01110*, 2020.

K. S. Walsh, D. P. McGovern, A. Clark, and R. G. O'Connell. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, 1464(1):242–268, Mar. 2020. doi: 10.1111/nyas.14321. URL `https://doi.org/10.1111/nyas.14321`.

H. Wang, T. Chen, S. Gui, T. Hu, J. Liu, and Z. Wang. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7449–7461. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/537d9b6c927223c796cac288cced29df-Paper.pdf`.

C. Wei and J. Z. Kolter. Certified robustness for deep equilibrium models via interval bound propagation. In *International Conference on Learning Representations*, 2021.

Z. Yang, T. Pang, and Y. Liu. A closer look at the adversarial robustness of deep equilibrium models. *Advances in Neural Information Processing Systems*, 35:10448–10461, 2022.

Z. Yang, P. Li, T. Pang, and Y. Liu. Improving adversarial robustness of deep equilibrium models with explicit regulations along the neural dynamics. 2023.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.