On the Trade-Off Between Transparency and Security in Adversarial Machine Learning

Extended Abstract

Lucas Fenaux
University of Waterloo
Waterloo, Canada
lucas.fenaux@uwaterloo.ca

Christopher Srinivasa
Borealis AI
Toronto, Canada
christopher.srinivasa@rbc.com

Florian Kerschbaum
University of Waterloo
Waterloo, Canada
florian.kerschbaum@uwaterloo.ca

ABSTRACT

Transparency and security are both central to Responsible AI, but they may conflict in adversarial settings. We investigate the strategic effect of transparency for agents through the lens of transferable adversarial example attacks. In transferable adversarial example attacks, attackers maliciously perturb their inputs using surrogate models to fool a defender's target model. These models can be defended or undefended, with both players having to decide which to use. Using a large-scale empirical evaluation of nine attacks across 181 models, we find that attackers are more successful when they match the defender's decision; hence, obscurity could be beneficial to the defender. With game theory, we analyze this trade-off between transparency and security by modeling this problem as both a Nash game and a Stackelberg game, and comparing the expected outcomes. Our analysis confirms that only knowing whether a defender's model is defended or not can sometimes be enough to damage its security. This result serves as an indicator of the general trade-off between transparency and security, suggesting that transparency in AI systems can be at odds with security. Beyond adversarial machine learning, our work illustrates how game-theoretic reasoning can uncover conflicts between transparency and security.

KEYWORDS

Adversarial Machine Learning; Game Theory; Responsible AI

ACM Reference Format:
Lucas Fenaux, Christopher Srinivasa, and Florian Kerschbaum. 2026. On the Trade-Off Between Transparency and Security in Adversarial Machine Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HCNI3628

1 INTRODUCTION

Responsible AI requires that systems and agents be both transparent and secure. These two qualities can intuitively appear at odds, especially in machine learning settings. An instance of this trade-off occurs in adversarial example attacks: a small perturbation added to an input that is imperceptible to humans but causes the AI model to misclassify it. The greater the attacker's access to the AI model it wants to attack, the better its attack will perform. We observe this principle in white-box attacks, which have complete access to the AI model, and black-box attacks, which can query the AI model, outperforming transferable attacks, which only have access to a surrogate model [12].

Another instance of the trade-off in adversarial example attacks, which has not been significantly studied, arises when deciding whether to disclose that a model is defended, without necessarily revealing the specific defense used. While it might seem trivial that this information would be useful to attackers, in the context of transferable adversarial example attacks, incorporating it into attacks is not straightforward and remains understudied [4, 10, 18]. We reinforce the finding that matching the target model's defense status meaningfully improves attack success rates. In particular, using defended surrogates to attack defended models is effective even when the defenses are not identical. Our findings are derived from a large-scale empirical evaluation in Section 2 and a game-theoretic analysis of the trade-off in Section 3, spanning nine transferable adversarial example attacks, 181 models, and two datasets.

Our game-theoretic analysis demonstrates that it can be beneficial to AI providers' security to conceal that they are defending against adversarial example attacks. Our results also support the notion that, when defenses are concealed, combining defenses can further enhance an AI system's robustness against adversarial attacks. Furthermore, we find that although many existing works benchmark transferable attacks using exclusively undefended surrogates, this approach can underestimate attack success rates by up to threefold. Our code implementation can be found at: https://github.com/LucasFenaux/transferable-attack-benchmark.

2 EMPIRICAL EVALUATION

We evaluate nine attacks: Admix [16], VNIFGSM [15], LGV [6], SSAH [9], BIA [17], OPS [7], PGN [5], CDTP [11], and AutoAttack [2] on the CIFAR10 [8] and ImageNet [3] datasets. To benchmark the attacks, we gather 92 undefended models and 89 defended models from the PyTorch-CIFAR GitHub repository [14], PyTorch's pre-trained models [13], and the RobustBench [1] framework. We measure attack success rate as the accuracy degradation incurred by attacks across the entire test set for CIFAR10, and the 5,000-image RobustBench subset of the ImageNet testing set. We present the accuracy degradation when using either an undefended or a defended surrogate to attack an undefended or a defended target model, averaged across the nine attacks in Table 1. We find that, on average, it is up to 3.2× more effective (in terms of increase in accuracy degradation) to use a defended surrogate than an undefended one to attack a defended model, while using a defended surrogate to attack an undefended model results in a 4.7× lower increase in accuracy degradation than using an undefended surrogate.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HCNI3628

Table 1: Mean accuracy degradation (in %) across nine attacks depending on the type of surrogate and target model. We include the increase compared to inferring without an attack in parentheses.

Target	Undefended		Defended	
Surrogate	Undefended	Defended	Undefended	Defended
CIFAR-10	**77.4 (+71.4)**	25.3 (+19.3)	14.8 (+2.5)	**20.4 (+8.1)**
ImageNet	**57.4 (+33.3)**	31.3 (+7.1)	30.0 (+1.7)	**32.3 (+4.0)**

3 TRADE-OFF ANALYSIS

Using the per-attack results from our empirical evaluation in Section 2, we can construct games to analyze the empirical advantage, for an attacker, of knowing whether the target model is defended. The first is a Nash game, which we call the Surrogate matching game. It represents the decision of whether to defend the target/surrogate model and its impact on attack success rate:

Players: The attacker is the row player and the defender is the column player. The attacker is associated with a fixed attack A.

Actions: The attacker chooses a distribution $\mathcal{S} \in \{\mathcal{U}, \mathcal{D}\}$ from the set containing the distribution of undefended models $\mathcal{U}$ and the distribution of defended models $\mathcal{D}$. The defender chooses a distribution $\mathcal{T} \in \{\mathcal{U}, \mathcal{D}\}$.

Utility functions: The utility function of the defender is the expected accuracy under attack of the target model's distribution:

$$u_d(\mathcal{S}, \mathcal{T}) = \mathbb{E}_{(S,T,x,y)\sim\mathcal{S}\times\mathcal{T}\times\mathcal{Z}}[\mathbb{1}(T(A(S,x,y)),y)] \quad (1)$$

Where $(x, y) \sim \mathcal{Z}$ denote input-label pairs. Since our game is zero-sum, the attacker's utility function is the accuracy degradation:

$$u_a(\mathcal{S}, \mathcal{T}) = 1 - u_d(\mathcal{S}, \mathcal{T}) \quad (2)$$

If, instead of deciding simultaneously, the attacker knows the defender's decision in advance, then the game behaves like a Stackelberg game with pure commitments. By measuring the difference in expected payoffs between the Nash and Stackelberg games, we can assess the advantage for the attacker of knowing the defender's decision (with only pure commitments, the advantage is non-negative). Empirically, we find that on CIFAR10, five of the nine attacks have a non-zero advantage, with an average advantage of 0.46% (in terms of accuracy degradation). Additionally, on ImageNet, six of the nine attacks have a non-zero advantage, with an average advantage of 0.33%.

In practice, however, an attacker would not be confined to a single attack method. Instead, they would select the attack with the best expected payoff among the available attacks and decide whether to defend their surrogate model. We name the game with this altered selection the Attack & Surrogate (A&S) game. When we factor in this selection, we find that the expected payoff for ImageNet differs by 0.18% between the Nash and Stackelberg games, with the attacker choosing the VNI-FGSM attack with a 95.82% probability and the OPS attack with a 4.18% probability. For CIFAR10, the expected payoff remains the same, as the attacker achieves the maximum expected payoff by always using the VNI-FGSM attack.

Finally, our results display that limiting surrogate selection to undefended surrogates when evaluating transferable attacks can underestimate their potency. To measure this underestimation concisely, we create another variant of our original game in which the attacker can only choose which attack to use and must use an undefended surrogate. We call this game the Attack game, and by comparing its expected payoff with that of the A&S game, we measure the underestimation of the potency of transferable attacks. We find a 3.73× (CIFAR10) and 2.15× (ImageNet) increase in mean accuracy degradation when the attacker can decide whether to defend its surrogate. Therefore, we encourage researchers who develop new transferable attacks or defenses against them to include defended surrogates in their evaluations.

4 CONCLUSION

In this work, we investigate the trade-off between transparency and security in adversarial machine learning through the lens of transferable adversarial example attacks. Our large-scale study on CIFAR10 and ImageNet showed that the attacker can benefit from matching the defender's defense decision. Thus, revealing even minimal information about a model's defense status can worsen its security. Our study also highlights an underestimation of the potency of transferable attacks against defended models by previous work.

We formalize defense decisions as Nash and Stackelberg games, depending on whether the defender is transparent. Then, we demonstrate that transparency, as prescribed by Responsible AI, can incur a security cost. This cost occurs even for something as minor as revealing that the target model is defended, without providing any additional information about the target model or the defense itself. While security through obscurity is often considered insufficient on its own, we show that, in the case of transferable adversarial example attacks, obscuring the defense status is a beneficial strategy for the defender. Finally, our results suggest that pure and overt strategies are unlikely to lead to optimal protection. Instead, they indicate that diversifying defenses as part of an overall defense strategy could enhance the defender's chances of success, especially when these defenses do not transfer well among themselves.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NSERC for grants RGPIN-2023-03244, IRC-537591, the Government of Ontario and the Royal Bank of Canada for funding this research.

REFERENCES

[1] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2020. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670* (2020).

[2] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*. PMLR, 2206–2216.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Mohamed Djilani, Salah Ghamizi, and Maxime Cordy. 2024. RobustBlack: Challenging Black-Box Adversarial Attacks on State-of-the-Art Defenses. *arXiv preprint arXiv:2412.20987* (2024).

[5] Zhijin Ge, Hongying Liu, Wang Xiaosen, Fanhua Shang, and Yuanyuan Liu. 2023. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems* 36 (2023), 70141–70161.

[6] Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. 2022. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *European Conference on Computer Vision*. Springer, 603–618.

[7] Yu Guo, Weiquan Liu, Qingshan Xu, Shijun Zheng, Shujun Huang, Yu Zang, Siqi Shen, Chenglu Wen, and Cheng Wang. 2025. Boosting Adversarial Transferability through Augmentation in Hypothesis Space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19175–19185.

[8] Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. In *Technical report*.

[9] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. 2022. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15315–15324.

[10] Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay. 2017. Cascade adversarial machine learning regularized with a unified embedding. *arXiv preprint arXiv:1708.02582* (2017).

[11] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. 2019. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems* 32 (2019).

[12] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 399–414.

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[14] Pytorch-cifar [n.d.]. Train CIFAR10 with PyTorch. https://github.com/kuangliu/pytorch-cifar

[15] Xiaosen Wang and Kun He. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1924–1933.

[16] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. 2021. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16158–16167.

[17] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. 2022. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. *arXiv preprint arXiv:2201.11528* (2022).

[18] Yechao Zhang, Shengshan Hu, Leo Yu Zhang, Junyu Shi, Minghui Li, Xiaogeng Liu, Wei Wan, and Hai Jin. 2024. Why does little robustness help? a further step towards understanding adversarial transferability. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3365–3384.