An Audit-based Approach for Model Trustworthiness:

Evaluation of Model Vulnerabilities to Adversarial Attack

Whitepaper by

Dr. Aravindhan Arunagiri, Senior Manager
Devendra Singh, Lead Data Scientist
Dr. Revendranath T, Project Manager

Mphasis
The Next Applied

Contents

1. Abstract

Machine Learning (ML) and Deep Learning (DL) models have made impressive progress in complex learning problems and revolutionized the field of artificial intelligence. However, the industrial adaptation of AI models in day-to-day applications is challenging due to security concerns related to adversarial attacks. Adversarial attacks are intended activities by scammers, cybercriminals, competitors and hackers to compromise the model performance, which includes accuracy, bias, fairness and/or exposing the sensitive data associated with them (for example training dataset, model inferences). The adversary attacks impact the model quality at various stages of its development life cycle, making it difficult to identify using existing evaluation approaches. Existing model evaluation approaches only identify the model performance issues occurring due to inefficiency in training data and model learning characteristics and cannot identify the vulnerabilities/behaviors of the attacks, trace their provenance and explain them to mitigate their risks.

Motivated by the limitations in existing evaluation models, in this article, we explain the impact of adversarial attacks on the AI/ML models and propose a Machine Learning (ML) audit-based approach to evaluate the model's robustness against adversarial attacks. This approach assesses the vulnerability of the model to different types of attacks and provides feedback on potential weaknesses. This feedback enables the model developers/deployers to identify and address potential security issues before the model is deployed in real-world scenarios.

Overall, this approach aims to provide a comprehensive solution to the security challenges faced by ML and DL models in the era of model vulnerabilities due to emerging trends in cyber-attacks and scams. Through an early detection of the vulnerabilities in the ML models, one can ensure reliability and trustworthiness in various downstream applications. The approach addresses the key concerns of ML systems to protect the data and internal behavior from various types of attacks. It discusses the necessary components and processes that are required to design a holistic and inclusive ML audit system to predict the model's vulnerability and required trustworthiness.

2. Introduction

State-of-the-art AI models use data-driven models for prediction and are trained over large and protected datasets with expensive processing capabilities and training time. In addition, rigorous testing is required before making the trained models available for production. The production-grade AI models make inferences on unknown/new/production data and are expected to deliver a high-quality predictions, safety and security. The whole life cycle of AI models is expensive. Yet, AI models made a remarkable progress in solving complex problems and production-ready off-the-shelf models are available through several cloud and individual offerings.

According to Fortune Business Insights, the global Deep Learning market size is projected to grow from $17.60 billion in 2023 to $188.58 billion by 2030, at a CAGR of 40.3% during the forecast period (Market Research Report, 2023). An increase in the usage of AI models in life-critical and real-world scenarios creates security and integrity concerns, especially in applications like self-driving cars, malware detection, drones, robotics, biometric authentication and automatic speech recognition, etc. (Dargan et al., 2020). Adversaries can manipulate legitimate inputs, in a manner imperceptible to humans, and cause trained models to produce incorrect outputs. Such attacks are recognized as adversarial attacks.

Szegedy et al. (2013) are pioneers in addressing the susceptibility of well-performing deep neural networks to adversarial attacks. Li et al. (2020) demonstrated the vulnerability of Deep Neural Network (DNN) in autonomous vehicles, where an adversary manipulated traffic signs to deceive the DNN predictions. Carlini and Farid (2020) investigated vulnerabilities of the trained forensic classifier (Deep Learning model) that distinguishes between real and synthetic images. The adversarial attacks due to synthetic images can impact the models validating images including digital signatures and biometric identifiers leading to high risks and costs.

AI models are vulnerable to adversarial attacks through the training and inference phases of the model development life cycle to the deployment phase when unseen data is passed to manipulate the model output. Multiple countermeasures have been proposed in recent years to mitigate the effects of adversarial attacks. Kurakin et al. (2016) proposed using adversarial training to protect the learner by augmenting the training set using both original and perturbed data. Open-source frameworks are found to be efficient in performing specific classes of attacks and identifying the affected model. Some open-source frameworks are developed to perform attacks on ML and DL models, access the robustness through techniques such as ex-Adversarial ML Threat Matrix, and tools like Counterfit, Adversarial-robustness-toolbox (Dickson, 2021).

The model robustness measures only explain the level of vulnerability of the model to attacks, in addition to other operational performance like prediction accuracy. To establish the trustworthiness of the AI models in real-world applications, it is important to understand the level of impact of adversarial attacks and the behavior of affected models. Therefore, tools to understand the level of robustness of the model to adversarial attacks are important and useful to the ecosystem. This whitepaper aims to propose the audit method to explain the level of robustness of the model to attacks and suggest their behavior in various deployment scenarios, enabling the developers/deployers to take early countermeasures to mitigate the concomitant loss.

The audit-based approach provides the user experts with insights into the model's immunity to scammer attacks and readiness to deployment.

This whitepaper is organized as follows: section 3 discusses [1] the impact of various types of adversarial attacks on model and data; [2] safety, security and privacy; [3] a set of metrics to identify and quantify the model vulnerability; and section 4 explains a comprehensive audit-based approach to evaluate various adversarial attacks on the model, and measure the vulnerability. The approach presents the user with the overall robustness of the model and its readiness for production and inferencing. The whitepaper details the benefits of an audit-based approach and the necessity of a comprehensive approach to measure the robustness of an ML model.

3. Adversarial Attacks

Scammers, cybercriminals, competitors and hackers deliberately exploit the vulnerabilities of AI models by making small, often (human) imperceptible changes to the input data and nudging the model to make incorrect predictions. The manipulated input data can be in the form of images, audio, text or any other type of data the model is trained to predict. This section discusses the common types of adversarial attacks with relevant examples.

Evasion attacks (Chakraborty et al., 2018) are the most prevalent in real-time situations such as attempts to evade the detection of malware and email spam. In an evasion attack, the attacker adds controlled perturbation to the data and manipulates the input data to deceive previously trained and deployed classifiers. During inferencing, the classifier misclassifies due to intended intrusion in the inference data. This attack degrades real-time predictions of the classifier although they are trained with legitimate training datasets.

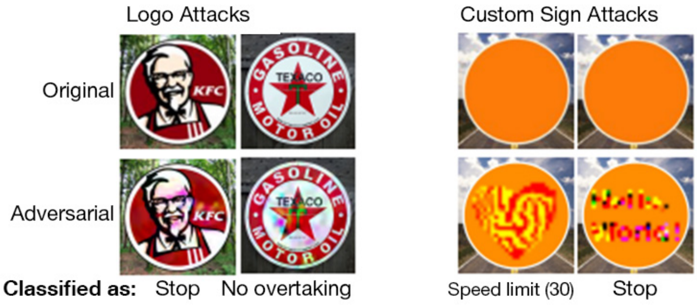

Figure 1: Shows the evasion attack on out-of-distribution examples for a traffic sign recognition ML pipeline, from Sevtlana et al.

The seriousness of evasion attacks is illustrated through two examples. First, an evasion attack leads to undesirable consequences on face authentication systems which are deployed in different settings to verify a person's identity. Assume that a facial authentication application is deployed to grant access to a secure building. An adversarial attack manipulates the input dataset in such a manner that it confuses the ML model to grant access to an attacker regardless of their identity. Such a compromised model-driven application may cause severe harm to the security of the building, resulting in a loss of trust and business.

Second, an evasion attack may wreck the entire navigation features and safety system of Autonomous Vehicles (AV). For example, road sign recognition model is a critical component of the AV system to safely navigate the roads. The vehicle's sensors and software use computer vision algorithms to identify road signs and interpret their meaning. An evasion attacker may modify the input data fed to road sign recognition models which may confuse a self-driving car to ignore a stop sign or misinterpret to exceed the speed limit. Such mispredictions could put passengers and other road users at potential risk of injury or death. Figure 1 shows an evasion attack on the ML model manipulating out-of-distribution samples and confusing the models to misinterpret the objects with high confidence.

Poisoning attack contaminates the training data or labels to alter the model's decision boundary to make the model underperform during deployment/inferencing. During the model life cycle, a model is trained with legitimate training data. After deployment, the model is regularly re-trained on freshly collected data, to keep it relevant. An attacker mostly poisons the model by injecting malicious samples in re-training data, which subsequently disrupts or influences re-training outcomes (best predictions). Figure 2 shows the impact of poisoning in the inference stage if the classifier model is (re)trained with perturbed data.

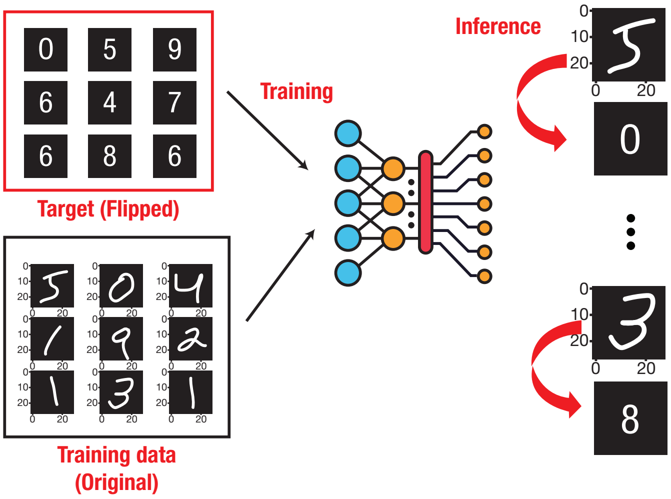

Figure 2: Poisoning attack example, showing the misclassification error by label-flipping from Liu et al., 2021

Privacy attacks are adversarial attacks with the primary goal of gaining knowledge from the model rather than deteriorating the model prediction performance. The privacy attackers intend to gain knowledge about: [1] the training data or properties of data, leading to data leakage and [2] model information to the extent of creating a model clone. The model information extraction enables an attacker to understand the internal workings of the model which can lead to manipulation of the model output. The privacy attacks are mainly of three types: membership inference, model inversion and model extraction.

1. *Membership Inference Attack* (MIA) tries to determine whether an input sample data has been used as part of the training data (Shokri et al., 2017). The MIA is further categorized into: (1) the White-Box attacks in which the attacker has access to model parameters, and (2) the Black-Box MIA in which the attacker has visibility only to the model output (predictions) and predicts for multiple instances. Therefore, it is important to secure access to the model parameters and gradients. Otherwise, effective White-Box membership inference attacks lead to high accuracy of predictions for the attackers.
2. *Model inversion attack* attempts to recover the private dataset used to train a supervised ML model. A successful model inversion attack generates realistic and diverse samples accurately describing each of the classes in the private dataset. Often, the attacker in this case uses a public dataset arbitrarily to target the train dataset (to be attacked).
3. *Model extraction* infers predictions of adversarial samples with the target ML model. Further, this sample and prediction are used to reconstruct/re-engineer a fresh ML model to mimic the behavior of the target ML model. For example, an extracted stock market prediction model can be used for financial benefit.

Table 1 summarizes the various types of adversarial attacks, implications on the model, data, and stages of different stages of the model development life cycle and useful metrics to interpret the robustness of the model towards a specific adversarial attack. Empirical robustness calculates the average minimum perturbation (to dataset) needed to create a successful attack. The successful attack is defined as the drop in the accuracy of the model (on perturbed data) by a predefined value. Loss sensitivity identifies the patterns in input data and allocates a high metric value for random data and a low value for real data. The CLEVER score measures the change in the gradient of loss between the actual training dataset and its adversarial samples and gives the level of memorization of the gradient loss (Weng et al. 2018).

Adversarial attack	ML life cycle stages impacted	Impacts	Metrics
Evasion	Inference phase	Model misclassification	Empirical robustness, Loss sensitivity, CLEVER score
Poisoning	Training phase	Model learning and alters the model decision boundary	Difference in performance of models trained on perturbed and pure data
Membership inference attack	Training data	Leaking of private information about training data	Accuracy of attack model
Model inversion	Training data	Adversary can generate training examples	Similarity of generated example and the true data point
Model extraction	Model	Stealing of the model, its behavior	Difference in accuracy of the subject model on same dataset

Table 1: Different adversarial attacks, their impact and metrics to quantify robustness

4. Audit-based Approach: Quantify the Robustness of the Model

The current AI model development life cycle is supported and accelerated using MLOps techniques and tools, thus reducing the lead time of production readiness of the models. In the AI life cycle, the input training data is a critical component influencing the performance of the model. After following the best practices to improve the quality of input data and transforming the data for the training phase, multi-step iterations of model training begin.

Model validation, retraining and testing may lead to an improvement in the performance of the model over unseen data. During this process, the existing training-validation frameworks make it difficult to decipher if the model performance degradation is due to data manipulation or poor learning during the training. The recent frameworks around explainability show a promise to identify the causal factors influencing the model performance due to the data patterns and model training parameters. The current training-validation frameworks cannot identify intended adversarial attacks and activate protection mechanisms to ensure the model's performance, safety and security. Furthermore, a model influenced by adversarial attack during the training passes the vulnerability to the entire AI model life cycle and downstream applications. Therefore, an audit-based approach to evaluate the model, identify adversarial attacks and identify the vulnerabilities is necessary in the AI model life cycle.

The proposed audit-based approach comprises capabilities to detect adversarial attack vulnerabilities of the AI model during various stages of its development life cycle. The approach comprises methods and tools to replicate cyber-attacks in a controlled fashion. A key component of the approach includes generating perturbed data, attacking or feeding the data to the prediction model and identifying the vulnerability of the model. The approach uses various measures that evaluate the level of model vulnerability and provides users with the measure of model robustness.

Figure 3 shows an approach comprising the activities required for auditing the model to identify its vulnerability to adversarial attacks. The adversarial attacks require perturbed data to perform the attack on the target model. The data synthesized with a known level of perturbation are fed to AI models to deceive or mislead from expected inferences. Techniques such as FGSM and PGD are used to create a controlled perturbation in the data.

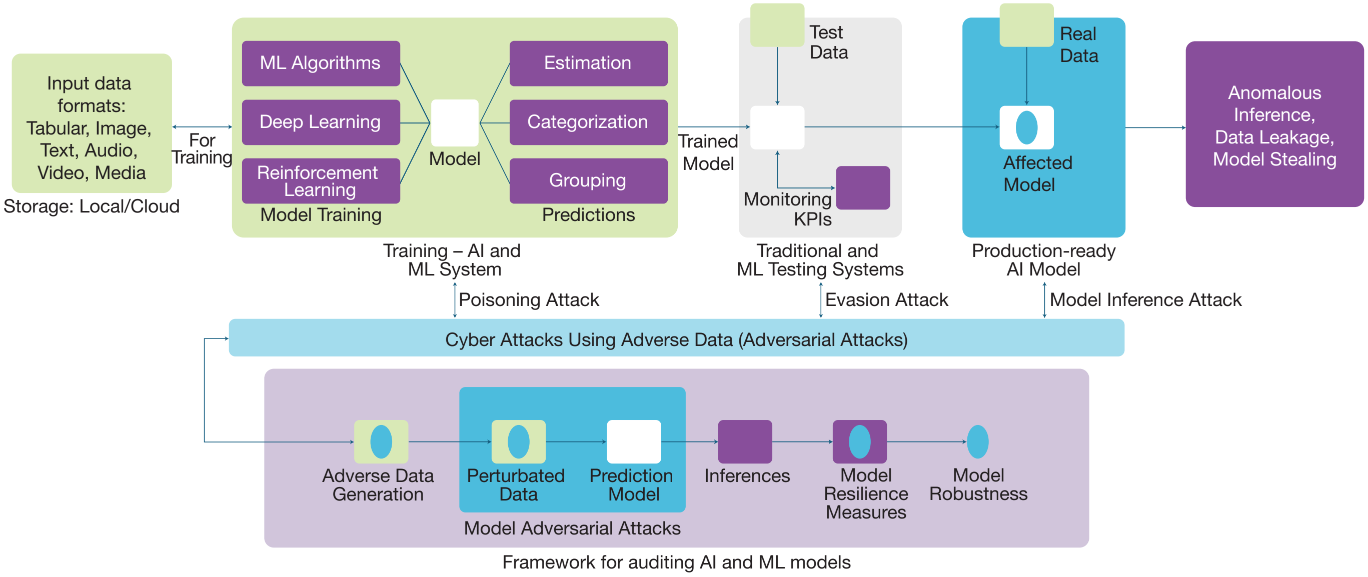
Input data formats: Tabular, Image, Text, Audio, Video, Media
Storage: Local/Cloud
For Training
ML Algorithms
Deep Learning
Reinforcement Learning
Model Training
Model
Estimation
Categorization
Grouping
Predictions
Training – AI and ML System
Trained Model
Test Data
Monitoring KPIs
Traditional and ML Testing Systems
Real Data
Affected Model
Production-ready AI Model
Anomalous Inference, Data Leakage, Model Stealing
Poisoning Attack
Evasion Attack
Model Inference Attack
Cyber Attacks Using Adverse Data (Adversarial Attacks)
Adverse Data Generation
Perturbated Data
Prediction Model
Model Adversarial Attacks
Inferences
Model Resilience Measures
Model Robustness
Framework for auditing AI and ML models

Figure 3: An audit-based approach for evaluating the adversarial attack on ML models

1. ***Fast Gradient Sign Method (FGSM)*** takes a single step in the direction of the sign of the gradient of the loss function with respect to the input for generating adversarial examples to fool the model.
2. ***Projected Gradient Descent (PGD)*** is an iterative optimization-based method for generating adversarial examples. It applies small perturbations to the input data over multiple iterations while constraining the perturbations to stay within a predefined epsilon bound. PGD is known for its robustness against various defense mechanisms.
3. ***DeepFool*** is an optimization-based technique with an aim to find the minimum adversarial perturbation required to push an input sample across the decision boundary of a model. It iteratively estimates the minimum perturbation based on the linear approximation of the decision boundary.

4.1 Model resilience measures

Table 1 shows a set of metrices to assess the robustness of the model against different kinds of adversarial attacks. The empirical robustness measures the minimum perturbation needed to perform a successful attack (Moosavi et al., 2016). CLEVER score (Weng et al., 2018) and the loss sensitivity (Arpit et al., 2017) quantify the robustness of the model. In addition to these metrices, sensitivity of model prediction performance to different levels of data perturbation can signify the model resilience. For example, in a poisoning attack, the difference in the accuracy of models trained on perturbed data and pure data can directly signify the model's robustness to a poisoning attack.

4.2 Model adversarial attack

The proposed audit-based approach explores vulnerabilities of the model to various adversarial attacks such as evasion attacks, poisoning attacks, model inversion attacks and membership inference attacks by simulating these attacks. Evasion attack evaluation involves replicating a controlled perturbation of data examples fed to the model and measuring the loss in the performance. Poisoning attack evaluations are performed by strategically modifying a subset of the training data or generating synthetic data samples to introduce biases in the model. To measure the robustness, a model is trained on both original data and back-door imputed data. The difference in the performance of both models shows the model's vulnerability. The model inversion attack is evaluated by querying the model multiple times to learn private information about the data used during the training process. The similarity between retrieved data and original data signifies the level of data leakage and hence the vulnerability of the model. Membership inference attacks aim to determine whether a specific data sample was part of the training dataset used to train the model. The model inversion attack evaluation is performed by training an attack model. The attack model's accuracy on validation data can signify the robustness of the model. This approach gives a versatile process to audit the model and explain the vulnerability with selected metrices.

5. Model Robustness Insights

Model robustness is a key aspect while building AI models to assure performance and reliability in diverse and challenging environments, ensuring practical applicability with security. The robustness against evasion attacks shows the model's resistance against manipulated inferencing data while the robustness against poisoning attacks signifies the model's ability to maintain performance even when trained on contaminated datasets. Membership Inference Attacks (MIA) robustness gauges the model's vulnerability to data leakage against the training data information. Model inversion robustness measure ensures the protection of sensitive information of training data by analyzing the model's outputs. In model extraction attacks, the robustness assesses the model's resilience to adversaries such as trying to create accurate surrogates or model clones based on the target model's response to queries.

6. Conclusion

The rapid growth in AI with the use of Deep Learning and Machine Learning models in life-critical decision-making led to the potential of adversarial attacks, creating vulnerabilities in models to compromise performance and data security. The proliferation of sophisticated attacks necessitates the assessment of the vulnerabilities of models to adversarial attacks before their deployment. The audit-based approach described in this paper enables complete evaluation and identification of model weaknesses at various stages of the Machine Learning pipeline. By proactively assessing and addressing these vulnerabilities, we can ensure the trustworthiness of Machine Learning systems and enhance protection against potential attacks.

7. References

1. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. CoRR abs/1312.6199 (2013). *http://arxiv.org/abs/1312.6199*
2. Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2016. Adversarial Machine Learning at Scale. CoRR abs/1611.01236 (2016). arXiv:1611.01236 *http://arxiv.org/abs/1611.01236*
3. Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In Security and Privacy (SP), 2017 IEEE Symposium on. IEEE, 3–18.
4. Ref: X. Liu, H. Li et al., “Privacy-Enhanced Federated Learning Against Poisoning Adversaries,” IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4574-4588, 2021.
5. Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 658-659).
6. Li, Y., Xu, X., Xiao, J., Li, S., & Shen, H. T. (2020). Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. IEEE Internet of Things Journal, 8(8), 6337-6347.
7. Svetlana Pavlitska and Nico Lambing and J. Marius Zöllner (2023). Adversarial Attacks on Traffic Sign Recognition: A Survey. In Proc. The international conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2023), July 2023, Tenerife, Canary Islands, Spain.
8. Tsui-Wei Weng and Huan Zhang and Pin-Yu Chen and Jinfeng Yi and Dong Su and Yupeng Gao and Cho-Jui Hsieh and Luca Daniel (2018), Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, In proc. Sixth International Conference on Learning Representations (ICLR 2018).
9. Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering, 27, 1071-1092.
10. Supreme Audit Institution (SAI) of Finland, Germany, the Netherlands, Norway, and the UK (2023). Auditing of Machine learning algorithms: A whitepaper for public auditors. The Office of the Auditor General of Norway.
11. M. Brundage et al. (2020): Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, *https://arxiv.org/abs/2004.07213*
12. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.
13. Deep Learning Market Growth, Share, (2023), Market research report, Fortune Business Insights.
14. Dickson, B. (2021, September 28). A developer’s Guide to Machine Learning Security | venturebeat. *https://venturebeat.com/business/a-developers-guide-to-machine-learning-security/*

About Mphasis

Mphasis’ purpose is to be the *“Driver in the Driverless Car”* for Global Enterprises by applying next-generation design, architecture and engineering services, to deliver scalable and sustainable software and technology solutions. Customer centricity is foundational to Mphasis, and is reflected in the Mphasis’ Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis’ Service Transformation approach helps ‘shrink the core’ through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis’ core reference architectures and tools, speed and innovation with domain expertise and specialization, combined with an integrated sustainability and purpose-led approach across its operations and solutions are key to building strong relationships with marquee clients. Click here to know more. (BSE: 526299; NSE: MPHASIS)

For more information, contact: marketinginfo.m@mphasis.com

USA
Mphasis Corporation
41 Madison Avenue
35th Floor, New York
New York 10010, USA
Tel: +1 (212) 686 6655

UK
Mphasis UK Limited
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Mphasis Limited
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

VAS 07/05/24 US LETTER BASIL 8955

www.mphasis.com