Human threats in AI-driven asset management systems: A case study on vibration-based bridge SHM

Y. Lan

Department of Civil Engineering, Aalto University, Espoo, Finland Department of Civil Engineering, University College Dublin, Belfield, Ireland

Z. Li, P. Acharya & W. Lin

Department of Civil Engineering, Aalto University, Espoo, Finland

ABSTRACT: As the world experiences its fourth industrial revolution, commonly known as Industry 4.0 (I4.0), practitioners are beginning to engage with I4.0 technologies, such as artificial intelligence (AI) and digital twins (DT). Such technologies are automating and digitizing previously analogue processes, enabling modern and sustainable asset management. Numerous studies have demonstrated the efficiency of those systems from monitoring to decision-making. However, as reliance on I4.0 technologies increases, a major concern arises: are these systems truly trustworthy? On one hand, there are worries about the system's accuracy and uncertainty, and on the other, the threat of malicious attacks from humans. For structural health monitoring (SHM), artificial disturbances can be introduced into an AI-based SHM system, perturbing healthy signals appear damaged, or vice versa, disguising damaged signals as healthy to obscure the true condition of a structure. Such malicious alteration could lead to catastrophic consequences. This paper discusses these by presenting a case study on traffic event-based deep-learning bridge SHMs, presenting methods for attacks and defense. The goal is to attract public attention to the vulnerability of AI-driven asset management systems and the development of defense means.

1 INTRODUCTION

Infrastructures are one of the major financial community assets and provide large benefits to society, yet their maintenance is still largely dependent on traditional methods (Malekjafarian et al. 2015). Inspectors conduct manual surveys to identify possible defects, after which asset managers organize necessary maintenance actions - a static and often reactive approach that results in delays, high labor demands, and significant cost inefficiencies (Dong & Catbas 2021). As the world enters the fourth industrial revolution, known as I4.0, new technological solutions are reshaping this landscape. Emerging techniques such as AI and DT are transforming infrastructure management by automating and digitizing tasks that were previously labor-intensive and analog (Cuellar et al. 2023). These technologies promise to enhance asset management by streamlining workflows and maximizing the efficient use of time, personnel, and resources. Leveraging I4.0 as a medium, real-time monitoring, predictive analytics, and data-driven decision-making enable proactive management strategies, reduce costs, and extend asset lifecycles (Opoku et al. 2021). Through such advancements, I4.0 offers a sustainable, efficient alternative to traditional practices, addressing the growing need for resilient infrastructure in a rapidly urbanizing world (Zhang et al. 2023).

Nonetheless, conservatives may question the extent to which AI, as a representative of I4.0, is trustworthy. This has driven the development of interpretable models and physics-informed

DOI: 10.1201/9781003595120-91

models, which are gaining increased attention (Xiong et al. 2023; Lan et al. 2024a, 2024b). Optimizing model accuracy while managing uncertainty is also a key area of research (Corbally & Malekjafarian 2023; Kamariotis et al. 2024; Lan et al. 2023a, 2023b). On the other hand, could AI potentially deceive decision-makers, or be exploited by malicious actors to do so? Concerns surrounding security, risk, ethics, and the need for defense and legislation within these systems are especially pressing. AI-driven asset management systems, in fact, may be vulnerable. Research in machine learning (ML) has shown that imperceptible perturbations targeting specific objectives can cause models, such as artificial neural networks (ANNs), to fail - this is known as adversarial attack (Kong et al. 2021). In autonomous driving, for instance, such attacks could lead to erroneous driving actions (e.g., due to erroneous traffic sign recognition), potentially causing traffic accidents (Eykholt et al. 2018). The same risks apply to infrastructure asset management; for example, a malicious classification of damaged structures as healthy poses serious risks. If undetected, such attacks could lead to a deterioration in structural health or even critical failures. Conversely, incorrectly labeling a healthy structure as damaged could lead to unnecessary maintenance and inspections, resulting in economic losses, and repeated misclassifications might undermine public trust in monitoring systems (Champneys et al. 2021). However, these issues currently receive little attention.

Adversarial attacks as human threats usually require the attacker to have full access to the system - such as knowledge of the model's architecture, parameters, gradient information, and access to training data, inputs, and outputs - constituting what is known as a white-box attack (Qiu et al. 2019). Given the complexity and accountability associated with asset management systems, one approach is to mandate transparency in their design, making white-box attacks within asset management systems somewhat feasible. On the other hand, attacks remain possible even when attackers only have access to the system's inputs and outputs, without any insight into the model's internal workings; this is referred to as a black-box attack (Papernot et al. 2017). In computer and data science, adversarial attacks often aim to make minimal changes to the model or input sample to produce significantly different outcomes. However, in practical asset management, such small modifications may be masked by inherent engineering uncertainties, and these modified samples are often challenging to introduce into real-world systems. This may contribute to the longstanding lack of attention from practitioners and researchers toward the human threats.

This paper demonstrates the human threats on an AI-driven bridge SHM asset management system, specifically targeting the deep learning (DL) network used in this study, from an attacker's perspective. Unlike computer science approaches, innovative strategies are proposed here to overcome the above engineering restrictions. These include a method for identifying and maximizing perturbations in features that have the greatest impact, along with an engineering-operable strategy that introduces perturbations physically through on-site shakers. Technical and legislative defense measures are then suggested. They are validated using a vibration database generated from a simulated vehicle-bridge interaction (VBI) model. The significance of this work lies not only in highlighting the vulnerabilities of AI-driven asset management systems but also in prompting further exploration of security, defense, legislation, and ethical considerations in the era of I4.0.

2 AI-DRIVEN BRIDGE SHM

2.1 SHM model and dataset

Consider an event-based AI-driven bridge SHM, specifically the structural responses caused by passing vehicles. It comprises three key components: data collection, model training, and condition assessment, as illustrated in Figure 1. In this work, data collection is conducted using mid-span sensors on a simply supported bridge, capturing vibrations that are then fed into a data preprocessing stage. Here, they are transformed into frequency domain data to create a set of vectors that serve as inputs to the SHM model. These vectors are stored in a dataset $\mathcal{D} = [d_1, d_2, d_3, \dots, d_{nt}]$, where nt denotes the number of traffic events. This study focuses on frequency domain signals, as structural condition changes can be more intuitively

discerned from frequency than from time-domain signals by human-beings. Model training and structural condition assessment are handled by a Convolutional Neural Network (CNN), a classic AI model with an architecture described in Table 1. During the condition assessment stage, data x from a new traffic event undergoes preprocessing and is then fed into the trained model \mathcal{M} . The model outputs a probability vector y, defined as:

$$y = \mathcal{M}(x) = \{P_1, P_2, P_3, \dots, P_C\}$$
 (1)

where each element P_i in vector y represents the probability that the input x belongs to the i-th label out of a total of C labels. The model's prediction is determined by the label corresponding to the highest probability in y. In this study, it is assumed that there are 10 structural states (C = 10).

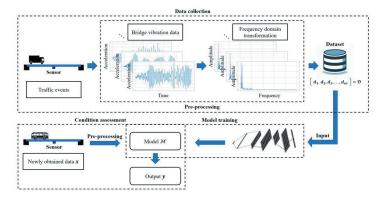


Figure 1. AI-driven SHM.

Table 1. CNN configurations.

Layer	Output shape	Parameter	Activation	
		Kernel number: 64; Kernel		
Convld	2500 × 64	size:10; Stride: 1; Padding: "same"	LeakyReLU	
Max pooling	1250×64	Kernel: 2; Stride: 2	None	
Convld	1250×128	Kernel: 128; Kernel size:10; Stride: 1; Padding: "same"	LeakyReLU	
Max pooling	625×128	Kernel: 2; Stride: 2	None	
Flatten	80000	None	None	
Dense	30	None	LeakyReLU	
Dense	10	None	Softmax	

This study utilizes a dataset of simulated bridge vibrations caused by traffic events (vehicle crossings). Data is generated using a simply supported Euler-Bernoulli (EB) beam model and a 2-DOF quarter-car model. Due to space limitations, specific VBI processes (e.g., model details, governing equations) are not discussed here; interested readers may refer to the authors' previous work (Lan et al. 2023; Li et al. 2023). The bridge parameters are as follows: mass per unit length $m = 2400 \, kg/m$, flexural rigidity $EI = 5.5 \times 10^9 \, N \cdot m^2$, length $L = 25 \, m$. The bridge is divided into 10 elements (n = 10), and damage location is represented by the element number, while damage severity is modeled as stiffness loss due to cracks or delamination, etc. The dataset aims to encompass a variety of traffic events. A vehicle crossing the bridge (e.g., a small bus or light truck) is characterized by the following parameters: $m_V^p = 1.28 \times 10^4 \, kg$, $m_V^p = 1.0 \times 10^3 \, kg$, $c_S^p = 1.0 \times 10^3 \, N \cdot s/m$, $c_V^p = 0$, $k_S^p = 4.0 \times 10^5 \, N/m$,

 $k_t^p = 3.5 \times 10^5 \, N/m$, and $v^p = 8 \, m/s$. Each passing vehicle follows a normal distribution corresponding to these baseline values. Each traffic event includes 5% environmental noise and random road roughness (Class A) to ensure the result robustness to these factors. The sampling rate is 1000 Hz, and for each different Damage Condition (DC), the dataset contains 1000 traffic events. Details are provided in Table 2. The data is then randomly split into a training set and a test set in a 9:1 ratio.

Table 2. Dataset.

DCs	Label	Description	Runs	DCs	Label	Description	Runs
DC 0	0	Healthy	1000	DC 5	5	1-st element, $\mu = 0.5$	1000
DC 1	1	5-th element, $\mu = 0.6$	1000	DC 6	6	3-rd element, $\mu = 0.5$	1000
DC 2	2	5-th element, $\mu = 0.7$	1000	DC 7	7	5-th element, $\mu = 0.5$	1000
DC 3	3	5-th element, $\mu = 0.8$	1000	DC 8	8	7-th element, $\mu = 0.5$	1000
DC 4	4	5-th element, $\mu = 0.9$	1000	DC 9	9	9-th element, $\mu = 0.5$	1000

2.2 Model performance

The CNN model is trained in a supervised manner, with a batch size of 32, Adam optimizer, learning rate of 1×10^{-4} , and cross-entropy loss function. The training history and model performance are illustrated in Figure 2 (200 epochs). The CNN achieved relatively high test accuracy (> 0.9), and there is no significant sign of overfitting. Due to the high similarity of signals, the model's suboptimal predictions mainly occur in distinguishing minor damage (DC 4) from its adjacent state (DC 3) and the healthy state (DC 0). Nevertheless, accuracy in these cases still exceeds 75%. For other DCs, model $\mathcal M$ performs quite well (91.2% accuracy). As an example, if decision-makers trust an AI-driven asset management system, they are more likely to make right decisions; this is why many studies focus on improving AI model accuracy. In fact, these models can be quite fragile, and most are exposed to human risks.

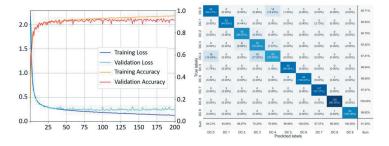


Figure 2. Model training history and performance.

3 HUMAN ATTACK

This paper introduces a black-box attack approach, implemented through system eavesdropping, feature perturbation, and sample insertion. It is a part of the FPA algorithm proposed by the authors (Lan et al. 2024). System eavesdropping involves monitoring each input signal and its output $y = \mathcal{M}(x)$, while recording the envelope of traffic events (i.e., the upper and lower response bounds of normal events). The first step in feature perturbation is to assess the contribution weight of each feature, bringing attention to the most impactful features. The algorithm, as previously proposed by the authors, initializes an empty set *Imp* for each sample. Within this loop, for each feature e_k in x_j , the algorithm perturbs e_k by adding Gaussian noise $\mathcal{N}(0,\mu^2)$ to yield a perturbed feature x_j [e_k]. It then calculates a new prediction y_j and expands Imp_j by the absolute difference between y_i and y_j . After iterating through

features in x_j , the algorithm merges Imp_j into the main set Imp. Once all samples have been processed, an average is calculated, Imp^{avg} ; the algorithm concludes by returning Imp^{avg} , with the top T most influential features assembled into an array, $p = \{f_1, f_2, f_3, \dots, f_T\}$.

The second step aims to maximize the model prediction error for p within the envelope bounds. One approach to do this is to use a Particle Swarm Optimization (PSO) optimizer; details on this algorithm can be found in reference (Poli et al. 2007). Here, a non-targeted attack strategy is introduced, which seeks only to ensure that the model does not produce correct results (i.e., tricking the SHM system into diagnosing a healthy bridge as damaged) to validate the approach. The objective function is defined as $F(x) = \mathcal{M}(x')_{Ori}$, where \mathcal{M} is the model, for $f_i \in p$, $x'[f_i] = x$, while all other components remain unchanged. *Ori* represents the original label. The physical constraint in this study is the observed traffic event envelope bounds. It is noteworthy that, since an on-site exciter will be used to physically insert these changes, negative vibration changes cannot be achieved. PSO seeks the optimal solution $x_{opt} = \{x'[f_1], x'[f_2], x'[f_i], \dots, x'[f_T]\}$ within these constraints to minimize the prediction probability of the original label:

$$\min_{\{f_i \in \mathbf{p}\}} P_{Ori} \Big(\mathcal{M}(\mathbf{x}_{\theta}) = \mathbf{y}_{\theta} | \mathbf{x}_{\theta}'[f_i] \Big) \tag{2}$$

Finally, an on-site exciter is employed to insert samples into the bridge vibration response. Since minor adversarial changes may be masked by the uncertainties in the physical process, we select the maximum permissible change within the physical constraints. An index vector $\zeta = \frac{x_{opt} - x_0[p]}{env_{max}[p] - x_0[p]}$ can be defined, where $\zeta = 1$ indicates the target frequencies of action for the exciter, denoted as $\mu = \{f_q | \zeta[f_q] = 1, f_q \in p\}$. The amplitude associated with μ is set as F. The exciter can then be controlled by Equation (3). Figure 3 summarizes the process of the human attack.

$$\mathbf{\Omega} = \mathbf{F}\cos(2\pi\mu t) \tag{3}$$

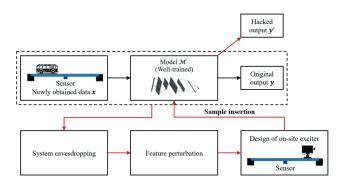


Figure 3. Process of the human attack.

4 THREATS AND DEFENSE

4.1 Threats

This section will explore the impact of human attacks on the proposed AI-driven SHM model, and the decisions based on it. Following this, defensive strategies will be suggested. By listening to 100 samples from DC 0, the envelope of normal vehicle passage events can be established, and the most influential features can be identified. For example, the top three are $p = \{32 \ Hz, 32.2 \ Hz, 31.8 \ Hz\}$. In fact, only a few features significantly impact the model's results; some align with physical insights, while others do not. This observation also raises concerns about model reliability and trustworthiness. However, due to space limits, they will not be the primary focus of this work.

It is found that, using the algorithm proposed in this study, only two feature alterations are sufficient to successfully attack the target model, and the spectrum after adversarial modification is very similar to the original. Minimal feature alterations in non-targeted attacks tend to shift predictions to the label most similar to the original (e.g., DC 4). Figure 4 presents a comparison of the spectra before and after modification alongside the model's predictions (10 features altered). When perturbing 10 features and considering all DCs, their confusion matrix is shown in Figure 5a. It can be observed that while DC 0 can be effectively disturbed to other DCs, some DCs remain resistant to such attacks due to distinct feature differences; altering only a few features is insufficient to bridge these gaps in feature space. However, by simply increasing the number of perturbed features, this robustness can be easily compromised (e.g., changing 100 features ensures a 100% attack success rate), as shown in Figure 5b. At this point, decision-makers may be misled in this example, leading to the belief that the structure is in an unhealthy state. This could result in further financial losses, such as unnecessary on-site inspections. This represents a human threat at the computer/data science level. However, in practical engineering, the insertion of such samples is often challenging. Due to space limitations of the conference paper, the robustness of different models under human threats and the success rates of attacks on various DCs have not been discussed here. Further exploration of them is encouraged in future studies.

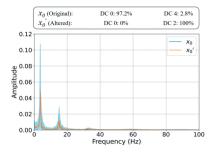


Figure 4. Adversarial results and performance comparison.

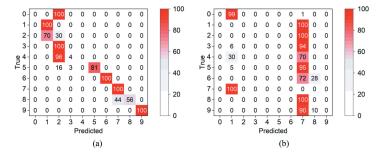


Figure 5. Adversarial confusion matrix: (a) 10 PFs, (b) 100 PFs.

We installed an on-site shaker on the bridge, which is controlled by Equation (3), to modify its vibration signals. Since this study primarily focuses on numerical simulations, the emphasis is on conceptual validation rather than immediate real-world implementation. In practical applications, an on-site shaker should be capable of reproducing a specified sinusoidal function with controllable frequency and amplitude. Theoretically, its placement on the bridge is not strictly constrained, as long as it can effectively introduce perturbations into the structural response. For this study, we assume the shaker is positioned at the mid-span of the bridge as a representative case. However, further experimental validation is required to confirm its feasibility in real-world scenarios. The samples x_{H}' (converted to the frequency domain) can be collected via bridge sensors. Figure 6a compares the sample x_{H}' with the adversarial sample x_{0}' obtained by the algorithm, showing a high degree of similarity. Figure 6b indicates that its

prediction aligns with the above adversarial results, where the model misclassifies the healthy sample DC 0 as DC 2 with nearly 100% probability. These findings demonstrate that the configured shaker can effectively and physically embed adversarial changes into the system; defense measures in data security are ineffective against such attacks.

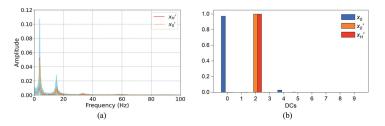


Figure 6. Samples from on-site shaker: (a) frequency domain, (b) prediction.

4.2 Defense

Based on the aforementioned methodology, defense measures can be proposed from both technical and legislative perspectives. Technically, defenses include developing interpretable AI models that provide insights into the contribution of each feature, their respective weights, and their physical significance. Features with unclear physical meaning or lacking robustness should be penalized. Additionally, employing multiple models for the same asset can enhance resilience, as adversarial attacks are often tailored to specific AI models; adversarial samples sensitive to one model may not affect another. Simultaneously deceiving multiple models significantly increases the technical difficulty of human attacks. Legislatively, clear legal provisions are needed, such as providing explicit legal definitions of human threats, determining whether accountability should be based on the consequences of the attack or the act itself, and addressing other legal ambiguities.

Due to space limitations, we can only present a few examples of potential measures. However, they provide valuable insights for future developments and may inspire further discussion.

5 CONCLUSION

This paper argues that AI-driven asset management systems are vulnerable and demonstrates how to hack a vibration-based bridge SHM from an attacker's perspective. This raises concerns about the interpretability and trustworthiness in I4.0 asset management, rather than just their efficiency, etc. But more than that, it explores the potential defense strategies against such attacks. They are demonstrated through a case of a typical traffic event-based bridge SHM. Based on the results, the following conclusions can be drawn:

- (1) A black-box human attack method, implemented through system eavesdropping, feature perturbation, and sample insertion, can hack AI asset management systems. Altering just two features is sufficient to untargetedly perturb the structural state from healthy to damaged.
- (2) For robust DCs, increasing the number of perturbed features improves the attack success rate. In all cases, the proposed method can achieve a 100% attack success rate by perturbing 100 features.
- (3) A customized shaker can physically and effectively embed adversarial changes into system samples; this extends defence work beyond the discipline of data security.
- (4) Several technical and legislative defence strategies are suggested, including interpretable AI frameworks, multiple-model approaches, and clear legal regulations. These measures collectively reduce risks from both technical and legal perspectives.

Future studies will involve configuring a real hacker exciter for experimentation, the key to which is the adaptive actuator as described in the text. Importantly, how to defend against such threats is a focus for future exploration. We hope this paper will attract valuable ideas and discussions.

ACKNOWLEDGMENT

This research is sponsored by the Jane and Aatos Erkko Foundation in Finland (Grant No. 210018).

REFERENCES

- Malekjafarian, A., McGetrick, P.J. & OBrien, E.J. 2015. A Review of Indirect Bridge Monitoring Using Passing Vehicles. *Shock and Vibration* 2015: 286139.
- Dong, C.Z. & Catbas, F.N. 2021. A review of computer vision–based structural health monitoring at local and global levels. *Structural Health Monitoring* 20(2): 692–743.
- Cuellar, S., Grisales, S. & Castaneda, D.I. 2023. Constructing tomorrow: A multifaceted exploration of Industry 4.0 scientific, patents, and market trend. *Automation in Construction* 156: 105113.
- Opoku, D.G.J., Perera, S., Osei-Kyei, R. & Rashidi, M. 2021. Digital twin application in the construction industry: A literature review. *Journal of Building Engineering* 40: 102726.
- Zhang, C., Wang, Z., Zhou, G., Chang, F., Ma, D., Jing, Y., Cheng, W., Ding, K. & Zhao, D. 2023. Towards new-generation human-centric smart manufacturing in Industry 5.0: A systematic review. Advanced Engineering Informatics 57: 102121.
- Xiong, J., Fink, O., Zhou, J. & Ma, Y. 2023. Controlled physics-informed data generation for deep learning-based remaining useful life prediction under unseen operation conditions. *Mechanical Systems and Signal Processing* 197: 110359.
- Lan, Y., Li, Z. & Lin, W. 2024a. Physics-guided diagnosis framework for bridge health monitoring using raw vehicle accelerations. *Mechanical Systems and Signal Processing* 206: 110899.
- Lan, Y., Li, Z. & Lin, W. 2024b. "Why Should I Trust You?": Exploring Interpretability in Machine Learning Approaches for Indirect SHM. Proceedings of the 11th European Workshop on Structural Health Monitoring (EWSHM 2024).
- Corbally, R. & Malekjafarian, A. 2023. A deep-learning framework for classifying the type, location, and severity of bridge damage using drive-by measurements. Computer-Aided Civil and Infrastructure Engineering 00: 1–20.
- Kamariotis, A., Chatzi, E., Straub, D., Dervilis, N., Goebel, K., Hughes, A.J., Lombaert, G., Papadimitriou, C., Papakonstantinou, K.G., Pozzi, M., Todd, M. & Worden, K. 2024. Monitoring-Supported Value Generation for Managing Structures and Infrastructure Systems. *Data-Centric Engineering* 5: e27.
- Lan, Y., Zhang, Y. & Lin, W. 2023a. Diagnosis algorithms for indirect bridge health monitoring via an optimized AdaBoost-linear SVM. *Engineering Structures* 275: 115239.
- Lan, Y., Li, Z. & Lin, W. 2023b. A Time-Domain Signal Processing Algorithm for Data-Driven Drive-by Inspection Methods: An Experimental Study. *Materials* 16(7): 2624.
- Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z. & Li, F. 2021. A Survey on Adversarial Attack in the Age of Artificial Intelligence. *Wireless Communications and Mobile Computing* 2021: e4907754.
- K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, & D. Song. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition: 1625–1634.
- Champneys, M.D., Green, A., Morales, J., Silva, M. & Mascarenas, D. 2021. On the vulnerability of data-driven structural health monitoring models to adversarial attack. *Structural Health Monitoring* 20(4): 1476–1493.
- Qiu, S., Liu, Q., Zhou, S. & Wu, C. 2019. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. *Applied Sciences* 9(5): 909.
- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B. & Swami, A. 2017. Practical Black-Box Attacks against Machine Learning. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security: 506–519.
- Lan, Y., Li, Z., Koski, K., Fülöp, L., Tirkkonen, T. & Lin, W. 2023. Bridge frequency identification in city bus monitoring: A coherence-PPI algorithm. *Engineering Structures* 296: 116913.
- Li, Z., Lan, Y. & Lin, W. 2023. Indirect damage detection for bridges using sensing and temporarily parked vehicles. *Engineering Structures* 291: 116459.
- Lan, Y., Li, Z. & Lin, W. 2024. Fooling Deep-Learning-Based Bridge Health Monitoring Models: A Hacker Vehicle. SSRN.
- Poli, R., Kennedy, J. & Blackwell, T. 2007. Particle swarm optimization. Swarm Intelligence 1(1): 33-57.