When Something Looks Too Good To Be True, It Usually Is! AI Is Causing A Credibility Crisis In Networking

Walter Willinger
NIKSUN, Inc., USA
wwillinger@niksun.com

Ronaldo A. Ferreira
UFMS, Brazil
raf@facom.ufms.br

Arpit Gupta
UCSB, USA
arpitgupta@ucsb.edu

Roman (Sylee) Beltiukov
UCSB, USA
rbeltiukov@ucsb.edu

Satyandra Guthula
UCSB, USA
satyandra@ucsb.edu

Lisandro Z. Granville
UFRGS, Brazil
granville@inf.ufrgs.br

Arthur S. Jacobs
UFRGS, Brazil
asjacobs@inf.ufrgs.br

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

The purpose of this editorial note is to raise awareness about a deeply concerning and yet much-overlooked development in the use of Artificial Intelligence (AI) and Machine Learning (ML) for solving problems in science in general and in networking in particular. To put it simply, in today's age of AI/ML, the much-publicized and well-documented "reproducibility crisis" in science is further compounded by an inconspicuous and rarely mentioned "credibility crisis." More to the point, by focusing on the area of networking research, we provide evidence that among the already small number of reproducible scientific publications that describe AI/ML-based solutions, even fewer, and often none, describe trained AI/ML models that are "credible;" that is, can be trusted to not only perform well in their original training domain but also in new and untested environments. We elaborate on the root cause of this credibility crisis, discuss why the credibility of AI/ML models is of paramount importance for their successful use in practice, and put forward an aggressive but imminently practical proposal for addressing this crisis head-on so as to pave the way for a future where networking research can reap the full benefits of AI/ML.

CCS CONCEPTS

• **Networks**; • **Computing methodologies** → **Machine learning**; **Artificial intelligence**;

KEYWORDS

Computer Networks, Artificial Intelligence, Machine Learning, Explainable AI, Credible AI Models

1 WHEN IT RAINS, IT POURS ...

For more than a decade, numerous scientific articles and headlines in the popular press have called attention to a deeply concerning "reproducibility crisis" in science in general [2] and in computer science in particular [1, 5]. Here, using the ACM-adopted terminology [10], a scientific paper is called *reproducible* if an independent researcher can replicate the results using the original author's artifacts (*e.g.,* code, data). Key among the voiced concerns is that this crisis undermines the integrity of reported scientific findings and raises questions about the reliability of existing scientific knowledge across various fields, thus threatening science as a whole. More recently, researchers have voiced their concerns about this reproducibility crisis spiraling out of control amidst an unprecedented surge of scientific publications driven by the rapid adoption and widespread use of artificial intelligence/machine learning (AI/ML) methods across all fields of science and engineering [17, 22, 24].

Taking this widely-documented reproducibility crisis in science in the age of AI/ML as a *fait accompli*, we report in this article on our experience scrutinizing recently published scientific publications in the networking area that leverage AI/ML-based methods to solve specific learning tasks of interest, are described as "success stories" that demonstrate the power of AI/ML, and that we have been able to reproduce. In short, our experience causes us to sound the alarm about a largely ignored problem that science is experiencing in the age of AI/ML and that we refer to as the "credibility crisis" of modern science. This credibility crisis on top of the generally acknowledged reproducibility crisis is succinctly captured by our main and unsettling finding: **of the few published AI/ML papers that are reproducible, even fewer, and often none, present trained AI/ML models that are "credible."** Here, we call a trained AI/ML model *credible* if it is *generalizable* (*i.e.,* performs well not only in the setting in which it was trained but also in new, unseen settings) and can be *trusted* (*i.e.,* we understand how the model makes its decisions and can predict when it is likely to fail or excel). Importantly, end users are reluctant, if not outright opposed, to relinquish control to trained models that are not credible, especially in cases where these models are intended to be used to make high-stakes decisions (*e.g.,* criminal justice system, operating production networks, controlling nuclear reactors).

Table 1: List of reproducible publications.

Research Paper	Learning Problem	Trained AI/ML model	Training Dataset	Journal/ Conference	Citiations (Google Scholar)
[36]	Detecting VPN traffic	1D-CNN	ISCXVPN2016 [11]	IEEE ISI'17	953
[30]	Detecting Heartbleed Attacks	RF Classifier	CIC-IDS-2017 [12]	IEEE ICISSP'18	4,209
[15]	Detecting Malicious Traffic	nPrintML	CIC-IDS-2017 [12]	ACM CCS'21	128
[15]	OS Finger-Printing	nPrintML	CIC-IDS-2017 [12]	ACM CCS'21	128
[37]	IOT Device Fingerprinting	IIsy (ML piece)	UNSW-IoT [31, 32]	HotNet'18	225
[21]	Anomaly Detection	Kitsune (KitNET)	Mirai trace [20]	NDSS'18	1,359
[19]	ABR Video Streaming	Pensieve	HSDPA Norway [26, 27]	SIGCOMM'17	1,453
[18]	Encrypted Traffic Classification	ET-BERT	Crossmarket [35]	WWW'22	221
[38]	Traffic Classification	YaTC	Crossmarket [35]	AAAI'23	32

2 SCOPE OF OUR EFFORT

To provide concrete evidence supporting this claimed credibility crisis, we focus on reproducible papers in the networking arena where the use of credible AI/ML models is critical. For one, being generalizable improves the models' applicability due to their ability to produce accurate predictions across a wide range of data, not just on the data they were trained on. Moreover, being trusted by network operators makes the models more dependable for use in real-world scenarios, especially for tasks that involve high-stakes decision-making such as correctly configuring networks and devices, avoiding collateral damage when controlling revenue-generating production traffic, effectively identifying and mitigating nefarious behavior, or limiting the damage caused by bad actors. We specifically examine reproducible networking research papers that claim "successful" applications of AI/ML (*i.e.,* as quantified by reported high F1 scores) to clearly defined problems in network security or performance.

Through an informal and ad-hoc survey of the relevant literature, we examined publications from various outlets. These ranged from high-impact conferences and journals to less competitive conference and workshop proceedings. We included papers with citation counts ranging from a few dozen to several thousand, published between 2017 and 2023. This survey led to the identification of eight papers describing a total of nine different trained AI/ML models, each developed for a specific learning problem and each resulting in F1 scores close to 1.[1] The models include Random Forest classifiers, 1D-CNN deep learning models, automated ML models, Reinforcement Learning models, and recently proposed network-specific foundation models. The learning problems that these models aim to solve span a range of concrete tasks, including (encrypted) traffic classification, OS and device fingerprinting, ABR video streaming, and intrusion detection. The papers are listed in Table 1, and for each of them, we include the considered learning problem, the utilized training data, the type of trained AI/ML model, the venue the original paper was published, and its citation count (Google Scholar, as of end of December 2024). Note that because of the problematic nature of the CIC-IDS-2017 dataset (see [16] for details), Table 1 does not include all published papers that describe AI/ML models that have used this dataset as training data. Doing so would simply boost the number of examples of published trained AI/ML models that lack credibility, irrespective of whether or not they are reproducible.

[1]Note that this yield is by and large consistent with recently reported numbers in [22] where the authors limited their search to papers from the four Tier 1 security conferences (*i.e.,* ACM-CCS, IEEE S&P, NDSS, and USENIX Security) that were published between 2013 and 2022.

3 BLAME THE STANDARD ML PIPELINE!

We focus on these eight papers to assess the rigor of the scientific methods employed in these reproducible studies and understand what these methods can tell us about the credibility of the proposed AI/ML models. To achieve this objective, we start by noting that all the trained models considered in these papers are black-box in nature and result from an application of the so-called *standard ML pipeline*. This pipeline details the widely-used workflow for developing AI/ML models that starts with describing the learning problem of interest, specifying a model architecture and a training dataset, and then combining these two ingredients to obtain a trained model whose performance is subsequently assessed with the exclusive help of an independent and identically distributed (IID) evaluation procedure. A hallmark of this procedure is that the trained model's expected predictive performance is evaluated on a test dataset that is an IID copy of the training dataset. In particular, it leaves the choice of what training data to use entirely up to the end users, whether they are domain experts or not, provides no guidance on assessing the suitability of the training data, and lacks the means to account for non-performance related model properties such as generalizability or trustworthiness.

In describing the approach we used to scrutinize each of the eight identified reproducible papers, we highlight two of its key aspects. First, our approach leverages recently reported findings that show that many modern ML pipelines, including the standard ML pipeline, suffer from *underspecification* and thus output trained AI/ML models that are not generalizable [7]. Here, an ML pipeline is said to be underspecified if it fails to specify a model in sufficient detail—an inability to capture essential structures that are expected to exist beyond the training domain. Importantly,

AI/ML models that are underspecified typically achieve their reported excellent performance by such means as identifying *shortcut learning strategies* (akin to "cheating"), exhibiting vulnerabilities to *out-of-distribution* samples of practical importance (akin to "rote learning"), or relying on *spurious correlations* (akin to using "lucky guesses"), all of which are clear examples that fully justify calling these models not credible.

The second key aspect of our methodology is the extensive use of Trustee, a novel global explainability tool we recently developed that allows end users to check if their trained black-box models that result from an application of the standard ML pipeline are underspecified and therefore not credible [16]. The main idea behind Trustee is to consider a trained AI/ML model (*i.e.,* the output of the standard ML pipeline) together with the dataset that was used to train this model and use these two ingredients as input to a process that extracts "white-box" models in the form of decision trees from the black-box input model, evaluates the fidelity of the extracted white-box models (with respect to the black-box input model), and outputs a decision tree that has high fidelity and a level of complexity (*i.e.,* tree size) that can be specified by the end-user. Trustee succeeds in this endeavor by applying a teacher-student dynamic derived from imitation learning to guide the training of decision trees that imitate the black-box model's decisions. While the high-fidelity requirement imposed on a Trustee-generated decision tree entails that, for all practical purposes, the obtained decision tree mimics the decision-making process of the black-box model, the user-specified level of complexity ensures that the size of the decision tree is manageable and therefore amenable to careful scrutiny by domain experts.

4 SUMMARY OF MAIN FINDINGS

While a detailed account of our efforts to examine the credibility of each of the nine different trained AI/ML models reported in the eight identified reproducible papers is beyond the scope of this article (but can be found in [14, 16]), the key lessons we learned are both telling and alarming and can be summarized as follows:

- Despite their reported excellent performance, none of the examined trained AI/ML models are credible.
- Each of the examined trained AI/ML models suffers from underspecification, and in each case, we can pinpoint the specific nature of the underspecification problem.
- The root cause of underspecification in all cases can be traced to the training data and specifically to problems with how the data was collected, generated, represented, or used.

5 CAVEATS

The popularity of the standard ML pipeline has arguably enabled transformational progress in many AI/ML application domains. However, our findings that the trained AI/ML models produced by this pipeline are typically not credible, and cannot be trusted, seriously question the use of this pipeline, at least for the AI/ML application domain of networking. In fact, in this application domain, our experience is in agreement with previous observations reported in [7] that the standard pipeline's indiscriminate use is the main reason why trained AI/ML models are more likely than not to suffer from underspecification and forces us to not only contest the pipeline's scientific value but also dispute its practical utility. At the same time, given the popularity of this pipeline among researchers across various disciplines in science and engineering, we fully expect that this credibility problem is not confined to the networking area but is a broader issue affecting other AI/ML application domains as well.

In addition to limiting the scope of our effort to the use of AI/ML for solving networking-specific problems, when using Trustee-extracted decision trees to examine the credibility of the output of the standard ML pipeline, our efforts have focused mainly on supervised learning problems (*e.g.,* classification, regression). In particular, we are not claiming that our approach can be used in any other AI application domain or is directly applicable to other problems such as unsupervised or self-supervised learning problems. In fact, we fully expect that even in the application domain of networking research, there exist learning problems that will pose challenges for using Trustee in its current form, and it will be informative to identify concrete use cases that stress-test Trustee's utility and/or suggest possible improvements to the current method [23]. Nevertheless, it is worth noting that we already had success in using Trustee-derived insights to show that the latest generation of transformer-based network foundation models is just as vulnerable to underspecification (and therefore lack credibility) as the more traditional AI/ML models produced by the standard ML pipeline [14].

With respect to our proposed approach that centers around using Trustee as the main vehicle for scrutinizing trained AI/ML models to see if they are underspecified and hence not credible, one obvious limitation is that the type of white-box model that Trustee extracts from a given trained black-box AI/ML model is limited to the class of decision-tree models. For one, although they are inherently interpretable, decision trees cannot be expected to adequately mimic the complex decision-making process of every possible black-box model. At the same time, decision trees that are trained from scratch (*e.g.,* using the well-known CART algorithm [4]) often struggle to achieve high accuracy while keeping the complexity (*i.e.,* tree size) in check. However, in this context, it is worth noting that Trustee-extracted decision trees are not decision trees that are trained from scratch. Instead, they are the result of applying a teacher-student dynamic derived from imitation learning. In effect, Trustee uses the trained black-box model to guide the training of a surrogate "white-box" model in the form of a decision tree that imitates the black-box's decisions. As a result, Trustee-extracted trees often avoid the accuracy-complexity tradeoff that decision trees typically face when trained from scratch.

6 TOWARDS CREDIBLE AI/ML MODELS

Unfortunately, spotting underspecification issues in trained AI/ML models is challenging. By its very nature, the standard ML pipeline focuses almost exclusively on demonstrating a trained AI/ML model's "effect" – asserting that it "works" in the sense that its expected predictive performance on the utilized test data is excellent (as measured, for example, by a high F1-score). However, by narrowly focusing on performance, this workflow is largely obscuring its innate inability to understand "cause"; that is, reasoning why a trained model works, how it works, and when (and why) it doesn't work. However, understanding cause is at the heart of examining

trained AI/ML models for their credibility and is key to relating properties of trained models, such as their generalizability to the quality of the utilized training data. In this sense, the standard ML pipeline exemplifies a major limitation of modern ML workflows, namely their inability to provide guidance on the choice of the "right" data to train AI/ML models that are both performant and generalizable.

Our approach to examine the credibility of trained AI/ML models that result from an application of the standard ML pipeline centers around the use of Trustee, which enables end users to consider their trained black-box models and trace the root causes of identified instances of underspecification back to issues with the utilized training data [16]. Based on our experience of using Trustee in the application domain of networking, the tool is particularly effective in illuminating both the generalizability problem of modern AI/ML that is at the heart of the credibility crisis reported in this article and the related but under-explored "garbage in, garbage out" (GIGO) problem of modern AI/ML, which refers to the basic notion that the quality of a trained AI/ML model directly depends on the quality of the training data. The eight identified reproducible papers that we examined are textbook examples of the prevalence of both of these problems in published AI/ML research.

With respect to the generalizability problem, Trustee is an improvement over the current generation of local explainability tools (*e.g.,* LIME [25], SHAP [29], LEMNA [13]) that have been developed to gain an understanding of how a trained black-box model makes individual decisions (or decisions in a local region around a particular input sample). While these tools can complement Trustee, they lack the capability of describing how a given black-box model makes decisions holistically. At the same time, Trustee has also led to advances in dealing with the GIGO problem of modern AI/ML. Specifically, it has recently been integrated into a new ML pipeline that is "closed-loop" in the sense that it facilitates the detection of instances of underspecification in trained AI/ML models, relates these instances to specific problems with the utilized training data, and suggests the collection of new training datasets that prevent the re-trained AI/ML models from exhibiting the same underspecification issues [3].

7 THE FUTURE OF SCIENTIFIC DISCOVERY

In limiting Trustee to only extracting decision-tree models, our approach has some similarity to [28], where the author argues against explaining black-box models as a separate "post-hoc" activity altogether and instead advocates using learning models such as decision trees or linear models that are inherently interpretable to start with. While we agree with much of the reasoning in [28], our use of Trustee to explain black-box models that have been trained to solve networking-specific tasks led us to take a more nuanced view on explaining black-box models. In particular, the networking domain is ripe with learning problems, where even domain experts cannot agree on what feature spaces to use for many of these learning problems. Clearly, this inability does not bode well for restricting model training from the start and argues for allowing a wider set of choices at the model specification stage of the standard ML pipeline to offer flexibility when it comes to feature engineering-related issues.

At the same time, our experience using Trustee to examine the credibility of trained black-box models to solve a given networking problem provides compelling evidence that Trustee-extracted decision trees can become useful and powerful vehicles for domain experts to check whether or not the original black-box model makes sensible decisions and can therefore be trusted or not. This then suggests tantalizing new opportunities for domain experts to potentially use Trustee-extracted decision trees in a "co-pilot" mode; that is, for the purpose of "learning from a trained AI/ML model," by which we mean the following two-step process. First, domain experts carefully inspect the decision tree that Trustee extracted from a given black-box model to see if it reveals decision-making strategies that initially strike them as questionable or even suspect. Second, upon encountering such strategies and painstakingly examining them, the domain experts conclude that they indeed represent legitimate, meaningful, and relevant strategies that have not been part of their current know-how but provide valuable new information and are fully deserving of being added to the experts' existing domain knowledge.

This combined use of a black-box model trained to solve a particular learning problem (*i.e.,* the output of the standard ML pipeline) and its corresponding high-fidelity white-box model (*i.e.,* the Trustee-extracted decision tree that mimics the decision-making process of the black-box model) has much in common with current discussions in the larger AI community. This ongoing discourse is largely triggered by the development of increasingly more advanced LLMs and application domain-specific foundation models [34]. However, rather than envisioning a future where AI solves particular problems that concern humankind and society as a whole, or proves specific conjectures in science and engineering in general or the fields of mathematics and computer science in particular, much of the ongoing discourse focuses on articulating a likely evolution towards an AI-assisted future. In this future, scientists of all ilk increasingly rely on AI to be used in the already alluded-to role as a co-pilot or, more specifically, leverage AI to create or hint at possible connections between different domains or topics, suggest new ideas, or serve as a conjecture generator [33]. Importantly, in this envisioned AI-assisted future, the role of the scientist shifts from actively formulating new ideas or conjectures or hypothesizing about possible connections to using human reasoning and existing domain knowledge to prove an AI-generated conjecture, assess the feasibility of an AI-formulated idea, or determine the validity of an AI-created connection [8].

Our proposed use of Trustee to explain trained black-box models can be viewed as a concrete instance of this envisioned AI-assisted future of scientific discovery in networking. On the one hand, the emergence of platforms such as AutoML [9] that fully automate the use of the standard ML pipeline are transforming the generation of trained AI/ML models for a given learning problem into a purely mechanical process and free the scientist from the mundane work required to perform the step-by-step instructions that comprise the standard ML pipeline. On the other hand, the responsibility of examining the credibility of a trained black-box model that this pipeline generated and scrutinizing it for possible underspecification issues rests squarely with the scientist who, for years to come, will have to deal with the paltry success rate that underspecified ML pipelines such as the standard ML pipeline have

for producing credible AI/ML models, irrespective of the learning problem at hand. In this context, Trustee becomes at once a "co-pilot" or a handy tool that allows the scientist to excel in what humans are inherently good at – making an informed decision as to whether or not the output produced by a modern ML pipeline can be considered credible.

8 TOWARDS AN AI-ASSISTED FUTURE OF NETWORKING RESEARCH: A PROPOSAL

Based on previous work [7] that provides an important account of the challenges that underspecification in modern ML pipelines poses for the scientific integrity of their output and based on our own experience using Trustee to scrutinize their output, we sound the alarm about a troubling credibility crisis that is affecting the rapidly growing scientific literature in the networking area in the age of AI/ML. The clearest manifestation of this crisis is that most trained AI/ML models described in these published works are not credible or generalizable and therefore cannot be trusted. To address this crisis caused by modern AI, we propose two unapologetically radical but imminently practical recommendations for the networking research community, in particular, and the science community, in general. Our proposal affects all aspects of research, from conducting scientific work to writing and reviewing scientific papers, and aims at realizing a future where our modern society can reap the full, science-backed benefits of AI.

First, to ensure that providing understanding remains a central purpose of science, researchers, in their role as "producers of AI-based science" (*e.g.,* authors), should agree that simply reporting on a trained AI/ML model that results from an application of the standard ML pipeline can no longer be considered a scientific achievement or research contribution. This is especially important given the emergence of platforms that automate this pipeline, transforming the creation of trained AI/ML models into a purely mechanical process. To be of actual scientific value, the development of trained AI/ML models for specific learning problems should lead to an "opening up" of science by making significant progress towards understanding "cause." In particular, to establish a backed-by-science AI, researchers should be required to address key questions such as why a proposed model works, how the trained model makes its decisions, and when the model does not work (and why not).

Second, given the growing awareness that modern ML pipelines, such as the popular standard ML pipeline, are highly vulnerable to underspecification, their almost exclusive use for developing trained AI/ML models for different learning problems is seriously compromised and defines an unacceptably low bar for declaring "success" in the sense of claiming that the models "work." Their use leads to a "dumbing down" of science by focusing almost solely on demonstrating "effect" while paying little to no attention to understanding "cause." To curtail their continued widespread use in their current form across the sciences, researchers, in their role as "consumers of AI-based science" (*e.g.,* reviewers), should be urged to outright reject submissions that merely report trained AI/ML models resulting from a straightforward application of the standard ML pipeline. Instead, future submissions should be required to demonstrate the credibility of trained models by ruling out underspecification as the primary reason for their reported excellent performance. At a minimum, it should be mandatory for future submissions to show that the trained AI/ML models they propose do not "cheat," do not engage in "rote learning," and do not rely on "lucky guesses."

A likely criticism of our proposal is that it asks for too much and puts the "burden of proof" squarely on the shoulders of the "producers of AI-based science." In particular, training AI/ML models that are credible and can be trusted will require researchers to spend considerably more time and effort when developing new AI/ML-based solutions, irrespective of the learning problem of interest. However, we argue that this role reversal is exactly what is needed to try and stem the current deluge of low-quality AI/ML papers in the sciences (*i.e.,* studies that focus exclusively on demonstrating "effect"), which is overwhelming the "consumers of AI-based science" (*e.g.,* artifact evaluation committees and reviewers) and is becoming unsustainable [6]. At the same time, we posit that our Trustee-based approach towards developing more credible AI/ML models is just a first step in equipping researchers with suitable tools to ease this burden and supplying them with better mechanisms that reduce their overheads and allows them to use their domain knowledge where most needed (*e.g.,* interpreting extracted decision trees and refining the collection of training data).

In conclusion, our proposal serves as an important reminder that a central and commonly agreed-upon purpose of science has been to provide understanding and expand human knowledge. We believe that this purpose remains true but attains new importance in the age of AI/ML. Our proposal appeals to researchers across science and engineering to take their responsibilities seriously and revive their pursuit of knowledge and understanding, and in doing so, they will stand a chance and succeed in preventing both the reproducibility and credibility crises in modern science from spiraling out of control.

ACKNOWLEDGMENTS

This work is partially funded by CNPq procs. 465446/2014-0, 142089/2018-4, 308101/2022-7, and 420934/2023-5; CAPES Finance Code 001; and FAPESP procs. 2020/05183-0, 2023/00811-0, and 2023/00812-7. Researchers at UCSB were supported by Cisco Research, and NSF Awards CNS-2323229, OAC-2126327, and OAC-2126281. Additionally, this research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award NERSC DDR-ERCAP0029768.

REFERENCES

[1] Vaibhav Bajpai, Anna Brunstrom, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, and Klaus Wehrle. 2019. The Dagstuhl beginners guide to reproducibility for experimental networking research. *ACM SIGCOMM Comput. Commun. Rev.* 49, 1 (2019), 24–30. https://dl.acm.org/doi/10.1145/3314212.3314217

[2] Monya Baker. 2016. 1,500 Scientists Lift the Lid on Reproducibility. *Nature* 533 (2016), 452–454. https://www.nature.com/articles/533452a#citeas

[3] Roman Beltiukov, Wenbo Guo, Arpit Gupta, and Walter Willinger. 2023. In Search of netUnicorn: A Data-Collection Platform to Develop Generalizable ML Models for Network Security Problems. In *Proc. ACM CCS'23.* ACM, Copenhagen, Denmark, 2217–2231. https://dl.acm.org/doi/10.1145/3576915.3623075

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA.

[5] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in Computer Systems Research. *Communications of ACM (CACM)* 59, 3 (2016), 62–69. https://dl.acm.org/doi/pdf/10.1145/2812803

[6] Jon Crowcroft. 2024. *10,000 maniacs and AI is destroying Computer Science, one topic at a time ...* http://paravirtualization.blogspot.com/2024/06/10000-manacs-and-ai-is-destroying.html

[7] Alexander D'Amour, Katherine Heller, Dan Moldovan, and et al. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61. https://www.jmlr.org/papers/volume23/20-1335/20-1335.pdf

[8] Christoph Drösser. 2024. AI Will Become Mathematicians' 'Co-Pilot'. *Scientific American* (June 2024). https://www.scientificamerican.com/article/ai-will-become-mathematicians-co-pilot/

[9] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. In *Proc. 7th ICML Workshop on Automated Machine Learning*. PMLR, Virtual, 16 pages. https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_7.pdf

[10] Association for Computing Machinery. 2020. *ACM Artifact Review and Badging Version 1.1, Aug. 24, 2020.* ACM. https://www.acm.org/publications/policies/artifact-review-and-badging-current

[11] Canadian Institute for Cybersecurity. 2016. *VPN-nonVPN dataset (ISCXVPN2016).* CIC. https://www.unb.ca/cic/datasets/vpn.html

[12] Canadian Institute for Cybersecurity. 2017. *Intrusion detection evaluation dataset (CIC-IDS2017).* CIC. https://www.unb.ca/cic/datasets/ids-2017.html

[13] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing. 2018. LEMNA: Explaining Deep Learning Based Security Applications. In *Proceedings of the ACM CCS'18*. ACM, Toronto, Canada, 364–379. https://doi.org/10.1145/3243734.3243792

[14] Satyandra Guthula, Navya Battula, Roman Beltiukov, Wenbo Guo, and Arpit Gupta. 2025. NetFound: Foundation Model for Network Security. arXiv:2310.17025v4 [cs.OS] https://arxiv.org/pdf/2310.17025v4

[15] J. Holland, P. Schmitt, N. Feamster, and P. Mittal. 2021. New Directions in Automated Traffic Analysis. In *Proc. of the 2021 ACM CCS'21*. ACM, Republic of Korea, 3366–3383. https://doi.org/10.1145/3460120.3484758

[16] Arthur S. Jacobs, Roman Beltiukov, Walter Willinger, Ronaldo A. Ferreira, Arpit Gupta, and Lisandro Z. Granville. 2022. AI/ML for Network Security: The Emperor has no Clothes. In *Proc. ACM CCS'22*. ACM, Los Angeles, CA, 1537–1551. https://doi.org/10.1145/3548606.3560609

[17] Charlotte Jee. 2019. Machine Learning is Contributing to a Reproducibility Crisis within Science . *MIT Technology Review* (February 2019), 1 pages. https://www.technologyreview.com/2019/02/18/137357/machine-learning-is-contributing-to-a-reproducibility-crisis-within-science

[18] Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, and Jing Yu. 2022. ET-BERT: A Contextualized Datagram Representation with Pre-training Transformers for Encrypted Traffic Classification. In *Proc. of the ACM Web Conference 2022 (WWW '22)*. ACM, Lyon, France, 633–642. https://doi.org/10.1145/3485447.3512217

[19] H. Mao, R. Netravali, and M. Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proc. of the ACM SIGCOMM'17*. ACM, Los Angeles, CA, 197–210. https://doi.org/10.1145/3098822.3098843 Code available at GitHub repository at https://github.com/hongzimao/pensieve..

[20] Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. 2018. *KitNET.* https://github.com/ymirsky/KitNET-py

[21] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai. 2018. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. In *Proc. of the 18th Network and Distributed System Security Symposium (NDSS'18)*. ISOC, San Diego, CA, 15 pages. https://doi.org/10.48550/ARXIV.1802.09089

[22] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. 2023. Get in Researchers; We're Measuring Reproducibility: A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In *Proc. ACM CCS'23*. ACM, Copenhagen, Denmark, 3433–3459. https://dl.acm.org/doi/pdf/10.1145/3576915.3623130

[23] Sagar Patel, Dongsu Han, Nina Narodystka, and Sangeetha Abdu Jyothi. 2024. Toward Trustworthy Learning-Enabled Systems with Concept-Based Explanations. In *Proc. 23rd ACM Workshop on Hot Topics in Networks (HotNets'24)*. ACM, Irvine, CA, 60–67. https://dl.acm.org/doi/abs/10.1145/3696348.3696894

[24] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning* 22, 164 (2021), 1–20. https://jmlr2020.csail.mit.edu/papers/volume22/20-303/20-303.pdf

[25] M. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. ACL, San Diego, CA, 97–101. https://doi.org/10.18653/v1/N16-3020

[26] Haakon Riiser et al. 2012. *HSDPA-bandwidth Logs for Mobile HTTP Streaming Scenarios.* UMASS. https://skulddata.cs.umass.edu/traces/mmsys/2013/pathbandwidth/

[27] H. Riiser, P. Vigmostad, C. Griwodz, and P. Halvorsen. 2013. Commute path bandwidth traces from 3G networks: analysis and applications. In *Proc. of the 4th ACM Multimedia Systems Conference*. ACM, Oslo, Norway, 114–118. https://dl.acm.org/doi/10.1145/2483977.2483991

[28] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (2019), 206–2015. https://www.nature.com/articles/s42256-019-0048-x

[29] L. S. Shapley. 2016. *17. A Value for n-Person Games.* Princeton University Press, Princeton, NJ, 307–318. https://doi.org/10.1515/9781400881970-018

[30] I. Sharafaldin., A. H. Lashkari., and A. A. Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proc. of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*. INSTICC, SciTePress, Funchal, Madeira - Portugal, 108–116. https://doi.org/10.5220/0006639801080116

[31] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman. 2019. Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics. *IEEE Transactions on Mobile Computing* 18, 8 (2019), 1745–1759. https://doi.org/10.1109/TMC.2018.2866249

[32] UNSW Sydney. 2018. *Data collected for our paper at IEEE TMC 2018.* UNSW. https://iotanalytics.unsw.edu.au/iottraces.html

[33] Terence Tao. 2023. *Embracing change and resetting expectations.* Microsoft. https://unlocked.microsoft.com/ai-anthology/terence-tao

[34] Microsoft Unlocked. 2023. *AI Anthology: A Collection of Essays on the Future of AI.* Microsoft. https://unlocked.microsoft.com/ai-anthology/

[35] Thijs van Ede, Riccardo Bortolameotti, Andrea Continella, Jingjing Ren, Daniel J. Dubois, Martina Lindorfer, David R. Choffnes, Maarten van Steen, and Andreas Peter. 2020. FlowPrint: Semi-Supervised Mobile-App Fingerprinting on Encrypted Network Traffic. In *Proc. of the 27th Network and Distributed System Security Symposium (NDSS'20)*. ISOC, San Diego, CA, 18 pages. https://github.com/Thijsvanede/FlowPrint/tree/master/datasets

[36] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang. 2017. End-to-end Encrypted Traffic Classification with One-dimensional Convolution Neural Networks. In *Proc. of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, Beijing, China, 43–48. https://doi.org/10.1109/ISI.2017.8004872 Code available at GitHub repository https://github.com/echowei/DeepTraffic..

[37] Z. Xiong and N. Zilberman. 2019. Do Switches Dream of Machine Learning? Toward In-Network Classification. In *Proc. of the 18th ACM HotNets'19*. ACM, Princeton, NJ, 25–33. https://doi.org/10.1145/3365609.3365864

[38] Ruijie Zhao, Mingwei Zhan, Xianwen Deng, Yanhao Wang, Yijun Wang, Guan Gui, and Zhi Xue. 2023. Yet Another Traffic Classifier: A Masked Autoencoder Based Traffic Transformer with Multi-level Flow Representation. In *Proc. of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI, Washington, DC, Article 605, 8 pages. https://doi.org/10.1609/aaai.v37i4.25674