On robust machine learning research projects

Dmitry Namiot
Lomonosov Moscow State University
Moscow, Russia
dnamiot@gmail.com

Evgeniy Ilyushin, Ivan Chizov
Lomonosov Moscow State University
Moscow, Russia
john.ilyushin@gmail.com, ichizhov@cs.msu.ru

***Abstract*— The increasing use of machine learning systems in critical applications (or attempts at such an application) naturally leads to the issues of robustness (reliability) of such systems. In fact, these questions are becoming the most important from the point of view of the practical application of machine learning systems. This is reflected, naturally, in a large number of works devoted to the assessment of the robustness of machine learning systems, the architecture of such systems, and the protection of machine learning systems. At the same time, the problems with robustness can arise both naturally, due to the different distribution of data at the training and practical stages, and as a result of targeted actions (attacks on machine learning systems). This article examines academic and industrial research projects focusing on robust machine learning models.**

I. INTRODUCTION

Machine learning systems have gained a lot of popularity recently. The realities of today are such that machine learning is used in all cases where there are no analytical models and algorithms for direct computation. At the same time, machine learning (deep learning) is today a practical synonym for the concept of artificial intelligence. Naturally, under such conditions, machine learning systems began to be used for critical operations. This is not necessarily related to military (special) applications. Control systems, autonomous vehicles, medical applications - there are already many examples of ML/DL (machine learning/deep learning) systems in critical applications.

Despite the impressive performance of DL algorithms, many recent studies raise concerns about the safety and reliability of machine learning models [1]. For example, Szegedy and others have demonstrated for the first time that DL models are severely vulnerable to carefully crafted adversarial (adversarial hereinafter - rebuttal) examples [2]. Similarly, various types of attacks (constructing adversarial examples) with poisoning (special modification) of data and models have been proposed against DL systems [3], and various methods of protection against such strategies have been proposed in the literature [4].

However, the reliability of the defense methods is also questionable, and various studies have shown that most defense methods are ineffective against a particular attack. It is the discovery that DL models are neither secure nor reliable that significantly impedes their practical deployment in security-critical applications such as healthcare forecasting, which is naturally vital.

Figure 1 shows the number of publications devoted to adversarial examples [5].

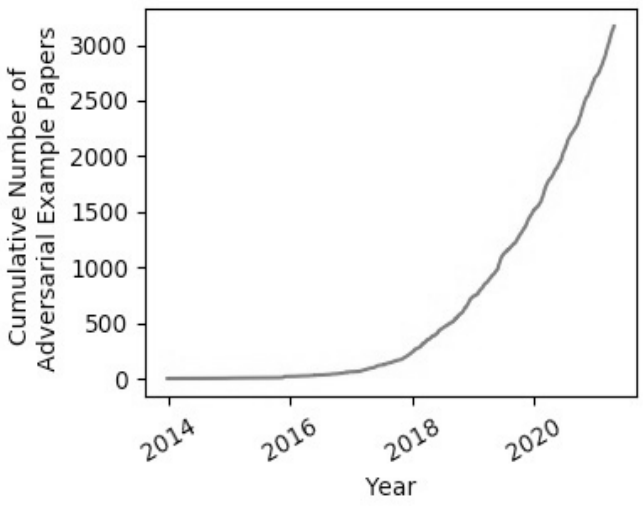

Fig. 1. Adversarial examples papers [5]

As you can see, the sharp rise began after 2018. Naturally, with such a fairly large number of publications, there are already many works that classify the existing problems with robustness (reliability), as well as present existing solutions. At the same time, the current state of this problem is such that there is no general solution that would guarantee the operation of an arbitrary machine learning system for all initial data. Would guarantee in the way it is done, for example, for software in avionics and other similar critical applications. This is discussed in more detail in section II.

Accordingly, the most important works in this direction have not yet been written (or, at least, have not been published publicly). Therefore, in our opinion, it would be interesting to consider existing academic and industrial projects in the field of robust machine learning, which are designed to somehow solve the existing problem.

This article was written within the framework of the project of the Department of Information Security of the Faculty of CMC, Lomonosov Moscow State University for the preparation of the master's program "Artificial Intelligence in Cybersecurity" [6].

The rest of the article is structured as follows. Section II summarizes the current state of research in robust machine learning models. Section III discusses industrial and academic projects in this area. Finally, Section IV is a conclusion.

II. ON THE CURRENT STATE OF ROBUST MACHINE LEARNING PROBLEM

Regardless of the models used and methods for obtaining independent parameters (features), the choice of analyzed variables, etc., any machine learning models always try to extend the results obtained during training to the entire general population of data. And in general, we have no reason to believe that this will be the case for all applications. It depends precisely on the field of application and our model, since this is a property of the general population of the data that we are researching. In other words, with only training data, it is impossible to say whether the conclusions drawn can be generalized. This is the main problem. Even a complete understanding of the principles of the system using training examples will not help if it turns out that the model does not work on real data. Accordingly, the problem of robustness (reliability) consists in checking (confirming) that the designed system can operate with data that differ from those on which it was trained.

Robust (reliable) and safe machine learning systems are systems whose behavior during operation does not differ from that declared at the testing and training stage.

Formally, for example, for a classification system, this can be defined as follows:

Some classifier C is δ - stable at a point only and if

$$\left\| \vec{X} - \vec{X}_0 \right\|_\infty \le \delta \Rightarrow C(\vec{X}) = C(\vec{X}_0) \qquad (1)$$

An intuitive definition says that if the difference between the original data in the feature space does not exceed δ, then such objects should be classified in the same way.

It is important that resilience refers only to the response to data changes and does not determine anything through the nature of these changes. Violation of this condition can be caused, for example, by the fact that the training dataset differs from the general population, there may be an incorrect choice of features, errors in the algorithm, as well as deliberate distortions in real data. In the latter case, they talk about attacks on machine learning systems. It is natural to assume that systems involved in critical applications are more likely to be attacked.

Strictly speaking, formula (1) determines the current state of the solution to the stability problem. Yes, many papers show that data modifications can violate the similarity condition of the classifier. But in full accordance with the definition, some minimal modifications are sought. Typical descriptions of attacks for machine learning systems that classify images - "changes imperceptible to the human eye allow you to bypass the limitations ...". Obviously, such a description clearly presupposes the presence of a person in the decision loop. But for automated systems, the size of the distortion is obviously irrelevant. Accordingly, attacks with an unlimited budget for changes can always be successful. Also in formula (1), we are talking about changes in the initial data of the model, but these initial data will not always (not in all models) be some direct characteristics of objects (such as, for example, individual points in images). In many cases, the features of models are artificial characteristics (for example, convolutions when working with sound). Changes (perturbations) of such parameters may not simply be associated with changes in real characteristics. The very idea that we consider a network as a black box precludes any proof of its properties, including robustness. And the requirement of explainability may conflict with the fact that we need more and more parameters to improve accuracy. The result is the lack of solutions to date that would guarantee stability for an arbitrary network on any data.

Briefly, the current state of the problem of creating robust machine learning models can be described by the expression “have understanding”. Yes, it is recognized that robustness (reliability) is today, perhaps, the main problem for the application of systems based on machine learning in critical areas. This is reflected in both academic articles and academic and industrial projects dedicated to robust machine learning.

As with any scientific field, it all starts with taxonomy. For natural reasons for the discrepancy between real data and those on which the model was trained, it is customary to talk about a distribution shift [7]. In the specified work [7], a shift is distinguished between a generalization of the subject area and a displacement of a subpopulation. When a generalization of the subject area is shifted, training and test distributions contain data from related, but different domains. For example, patient records but obtained from different hospitals or images were taken with different cameras, etc. In a subset shift, the training and test data are different subsets of the same distribution. One of the most common (and most studied) is the so-called covariance shift [8]. Here we assume that although the distribution of the input data may change over time, the labeling function, that is, the conditional distribution $P(y \mid x)$, does not change (x here presents our input, y – output). That is, the problem arises from a shift in the distribution of covariates. The shift of the label is the opposite situation, the probabilities for the labels (conclusions) $P(y)$ change, and the conditional probabilities $P(x \mid y)$ remain constant. For example, a medical diagnostic system, where the probability of meeting a diagnosis y decreases over time, and the classifier for making a diagnosis $P(x \mid y)$ does not change. And finally, concept shift is when the definitions of labels (conclusions) themselves change. For the same example of a diagnostic system, this is a change in the criteria for making diagnoses (conclusions).

Artificially created problems for machine learning systems are called attacks. It also has its own classification (and there are many such classifications).

For example, the following table provides examples of classifying attacks (and defense methods) depending on the attacked component of a machine learning system.

Table 1. Some examples of attacks and defenses

Attack	Place (stage) of application	Targeted parameters	Countermeasures (defenses)
Adversarial Attack	Usage (deployment)	Input data	Gradient Masking, Pre-Processing Filters, Adversarial Retraining
Backdoor Attack	Training	Model (network) parameters	Pruning, Fine Tuning
Data poisoning	Training, Usage (deployment)	Input data	Encryption, Local Training
IP stealing	Usage (deployment)	Output data	Obfuscation, Encryption
Neural-level trojan	Training	Output data	Data filtering
Side-channel Attack	Usage (deployment)	Output data	Randomness

And these, of course, are just a few examples of several hundred existing attacks, presented for illustrative purposes only. It is the description of such attacks, countermeasures and the construction of tables like the one above that hundreds of works are devoted to. All development in this area follows the following scheme: a description of a new attack (refuting examples) for a machine learning model - building a defense (more often - restrictions for attackers) - a new attack, etc.

At the same time, the problems identified in the classification of attacks should not be considered so rare. In fact, the same poisoning attacks happen much more often than one might think. A simple example - social networks can use (actually use) additional training of their recommendation systems based on real user behavior. And the "desired" behavior can be easily modeled. Another example related to dataset poisoning is simply errors in the markup. Many well-known datasets that are used in different systems, and are also the basis for pretrained networks, simply contain a lot of errors. An example is the work of MIT [9].

The current state of the so-called adversarial machine learning can be compactly illustrated in Figure 2.

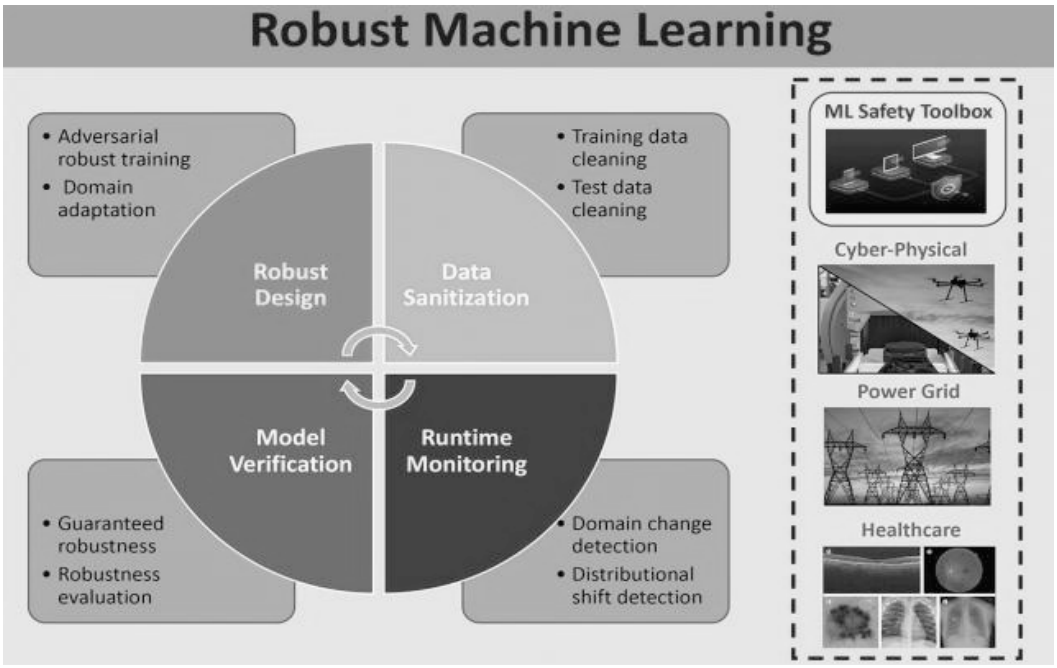

Fig. 2. Robust machine learning [10]

At the same time, only the verification of models is a real confirmation of robustness. The definition of a shift is just a definition of a change in the nature of the data. Data cleaning and filtering is only a potential opportunity to restrict malicious data where possible. For example, when working with voice data, we can try to limit the possible frequency range. Adversarial training, in short, is the inclusion of possible offensive data in training sets.

Verification as a direction naturally attracts a lot of attention, but today there are no examples of using this approach for real (large) neural networks [11, 12].

Existing reviews in this area, for example, [13], naturally describe the current state of the problem and existing (published) works. And the idea of this article is an assessment of the direction of future work.

III. Academic and Industrial Projects Dedicated to Researching Robust Machine Learning

The state order for this kind of work is obvious. The Joint Artificial Intelligence Center, created by the Pentagon to help the U.S. military use AI, recently formed a unit to collect, validate, and distribute open source machine learning and industry models to groups in the Department of Defense. Part of this effort points to a key issue with the use of AI for military purposes. The machine learning red team, known as the Test and Evaluation Team, will examine pre-trained models for weaknesses. Another cybersecurity group examines AI code and data for hidden vulnerabilities [14].

The rationale from the verbatim quote: "We don't know how to build systems that are completely resistant to adversarial attacks." The content of the work is, of course, unknown, but, for example, there is an assessment of the state of the problem and a work plan in the report of the Center for Security and New Technologies at Georgetown University, which is involved in these works. In particular, the report says that "data poisoning" in Artificial Intelligence (AI) systems can pose a serious threat to national security. This would involve infiltrating the process used to train the AI model, perhaps by using a volunteer agent to tag images fed to the algorithm, or by posting images on a network that are collected and fed into the AI model [15].

The report deals with the organization of work to protect shared (reusable) resources in AI systems, which include datasets, pre-trained models, and development tools. The author of the report is Andrew Lohn of RAND Corporation, according to his Google Scholar profile. Recently, there is a brief report on adversarial attacks and associated risks [16] with a rather disappointing conclusion: “The proposed protection methods can only provide a short-term advantage. Attacker-defender confrontation in machine learning systems is like a cat-and-mouse game. At the same time, the defenders lose, their defense methods are bypassed (overcome), and they do not yet keep up with the attackers. However, defensive measures can increase costs for attackers in some narrow cases, and a proper understanding of machine learning vulnerabilities can help defenders mitigate the risk. It can be expected that the effectiveness of defensive strategies and tactics will change over the years, but still will not be able to withstand more complex attacks”.

Another example is the DARPA Guaranteeing AI Robustness Against Deception (GARD) program [17]. GARD is committed to providing the theoretical foundations of a machine learning system to identify system vulnerabilities, characterize properties that will improve system reliability, and encourage the creation of effective defenses. Currently, protection against machine learning is usually very specific and only effective against certain attacks. GARD is committed to developing defenses that can withstand a wide range of attacks. In addition, current paradigms for assessing AI resilience often focus on simplistic measures that may not be related to security. To test for safety relevance and broad applicability, GARD safeguards will be measured on a new testbed using scenario-based assessments.

The latter seems to be especially interesting in terms of work. It has already been noted above that at present, when building protection systems (when assessing stability), rather artificial data perturbations are used, which proceed from the invisibility of changes for a person. At the same time, it is obvious that a person is not necessarily present in the decision-making chain for all applications. And for critical applications, it is completely absent. It is possible that scenarios will be associated with more realistic changes.

To carry out the work, 17 organizations were selected, including the universities of MIT, Carnegie Mellon, as well as Intel, and IBM [18].

The US federal government is funding a national artificial intelligence program. Within the framework of this program, 7 institutes were selected in different directions, which receive federal funding for researching various aspects of artificial intelligence. The foundations (basic elements) of machine learning systems within the framework of this program are dealt with by a specially created Institute of Machine Learning at the University of Texas [19].

Robust machine learning is one of his main research areas: “Teaching machine learning models is computationally intensive, and deciding how to set parameters is more art than science. There is no good mathematical basis for how to set scales and levers. The second direction is resistance to both data errors and deliberate malicious manipulations. The current model, based on a huge amount of good quality (valid) training data, is unsuitable for most applications. As people try to use machine learning models for high-stakes applications, this problem will be exacerbated by hacking attempts. It is the development of robust machine learning models that will be the key to their widespread adoption” [20].

It is noted that a very important part of the development of large-scale artificial intelligence or machine learning systems is the quantification and assessment of uncertainty, because without this it is possible, in fact, to make catastrophic decisions in the era of big data.

At the same time, we have not yet found a large number of publications from this group of researchers on the stated topics. Either there are no results yet, or not everything is published yet.

The Information Science and Technology Institute (ISTI), a national laboratory in Los Alamos and its National Security Education Center [21], has declared the interpretability and explainability of machine learning models as the first priority for their research: “In the field of national security, there is an urgent need for interpretable machine learning models, especially, in critical (high-risk) applications. Unfortunately, most of the generally accepted interpretability methods are focused specifically on image processing and classification models trained on labeled data. Meanwhile, many applications of national importance in Los Alamos and elsewhere use time

series, text, or numbers in addition to images. However, they require the use of unsupervised models for tasks such as knowledge extraction or anomaly detection. In the summer of 2021, projects under this focal area will include the development and / or evaluation of interpretability methods for models that use text and / or time series for national security applications "[22].

At another national laboratory (Livermore), the Center for Applied Computing has two projects in our area of interest - robust machine learning [23] and explainable artificial intelligence systems [24].

Uncertainty quantification and interpretable ML are critical to creating trust and enabling users to gain insights into models and data. This is a very important point - we cannot verify (prove) anything for the black box. Black box models are not testable (and therefore cannot be declared robust) by definition.

Uncertainty quantification (UQ) techniques play a key role in reducing the impact of uncertainties, both during optimization processes and in decision making. They have been used to solve many real-life problems in science and technology. Bayesian approximation and ensemble learning methods are two commonly used types of uncertainty quantification (UQ) methods. Various UQ techniques are used in applications such as computer vision (such as self-driving cars and object detection), image processing (such as image restoration), medical image analysis (such as classification and segmentation of medical images), natural language processing (such as classification texts, texts on social networks and risk assessment of recidivism), bioinformatics, etc. [26]

Figure 3 from the Center's poster illustrates the area of Security & Privacy.

Certified model training guarantees the absence of trojans and bookmarks. And the importance of model verification was noted earlier.

The Intelligence Advanced Research Projects Activity (IARPA) [27] invests in high-risk and high-paying research programs to address some of the most challenging challenges facing agencies and disciplines in the intelligence community. Among the list of projects is, for example, a research project on bookmarks in machine learning systems [28]. Publications on this project are available [29]; the work was carried out at the University of Berkeley.

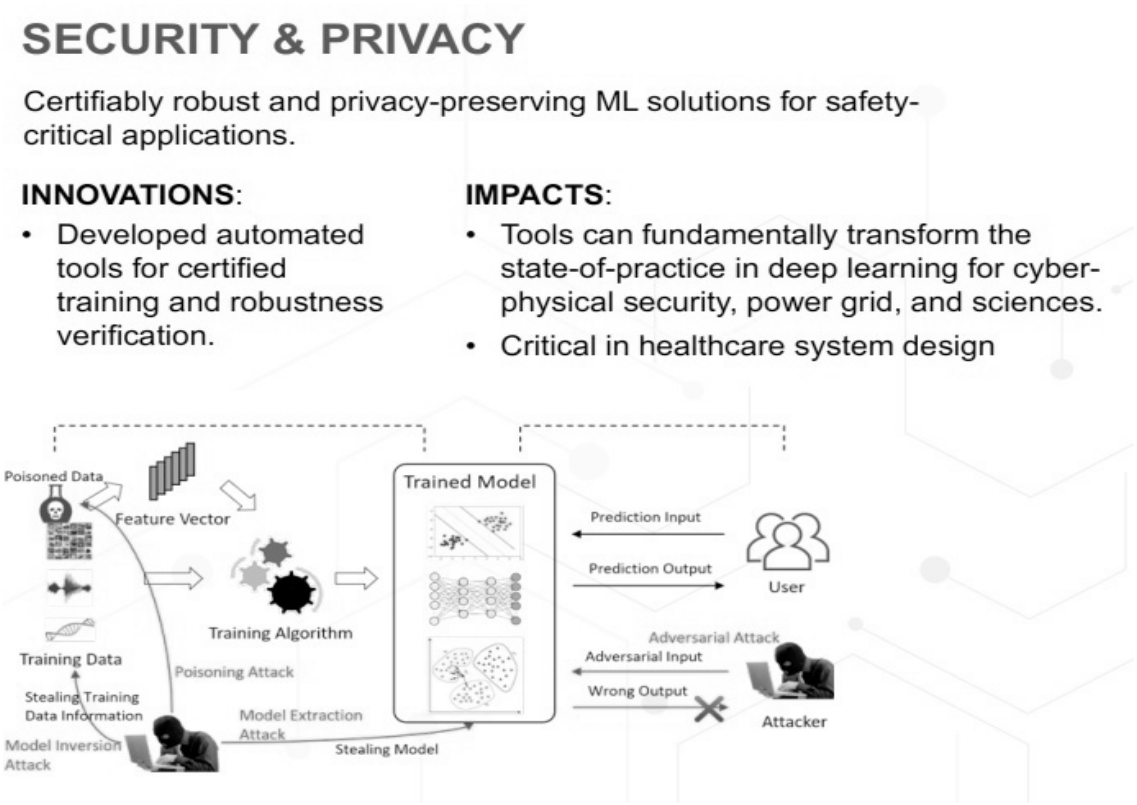

Fig. 3. Security and privacy at Livermore [25]

At the European level, one can note the ELLIS project - European Laboratory for Machine Learning [30]. Among his programs [31] are the direction of robust machine learning [32], as well as interpreted machine learning [33]

However, Google Scholar sees only a small number of publications mentioning ellis.eu. The articles discuss why the European community lags behind China in matters of artificial intelligence [34].

Among the academic and industrial projects, we would first of all note the group at MIT CSAIL [35], led by Alexander Madry [36] and the Center for Deployable Machine Learning (CDML) [37]. Links to the papers published on the laboratory's website [38] include the source code of the Python

library for robustness testing [39]. In particular, this group created a textbook on Adversarial machine learning [40].

In terms of work, in general, this can be described as adversarial training. At the same time, given the large composition of the groups, they cover a fairly wide range of problems and have an impressive number of publications, as well as very interesting PhD dissertations.

MIT employees are also behind the Robust ML open source security repository [41]. The site publishes protections against white-box attacks on image classification systems (here the authors follow the general trend when everything related to robustness refers, in the overwhelming majority of cases, to working with images). Refutations of defenses (that is, attacks that overcome them) are also published here. It should be understood that protection, in this case, is, for example, guaranteed boundaries for the loss function for a given restriction on modification [42]. That is, this is in no way a guaranteed confirmation of operability, as it is commonly understood, for example, for software of critical systems.

At the same level of importance, if not higher, should be considered Google (Deepmind). Their own manifesto for robust machine learning [43] notes that from a programmer's point of view, a bug is any behavior that is incompatible with a specification, that is, the intended functionality of a system. Deepmind conducts research into methods for assessing the conformity of machine learning systems not only to the training and testing set, but also to the list of specifications describing the desired properties of the system. Such properties may include resilience to small enough input disturbances, safety constraints to avoid catastrophic failures, or making predictions consistent with the laws of physics.

As part of such a study, Deepmind identifies three important technical challenges that, according to the company, will have to be solved by the machine learning community (that is, the tasks are completely general in nature):

1) Effective testing of compliance with specifications. Deepmind is exploring effective ways to validate machine learning systems meet the properties (such as invariance or reliability) desired by the system designer and users. One approach to identifying cases where a model may not fit the desired behavior is to systematically look for the worst results during the assessment.

2) Train (train) machine learning models according to specifications. Even with a large amount of training data, standard machine learning algorithms can create predictive models that make predictions inconsistent with desired specifications, such as reliability or fairness. Accordingly, this requires a revision of the training algorithms so that models can be created that not only fit the training data well, but are also consistent with the given specifications.

3) Formal proof of compliance of machine learning models with specifications. There is a need for algorithms that can verify that the model's predictions are provably consistent with the specification of interest for all possible inputs. Although such algorithms have been studied in the field of formal validation for several decades, these approaches are difficult to scale to modern deep learning systems.

The following problems are stated as specific directions:

a) Learning adversarial evaluation and validation: As AI systems scale and become more complex, it will become increasingly difficult to develop adversarial evaluation and validation algorithms that are well adapted to the AI model. If we can harness the power of AI to facilitate assessment and validation, the process can be scaled up.

b) Developing publicly available adversarial assessment and validation tools: It is important to provide AI engineers and practitioners with easy-to-use tools that predictably (before negative consequences occur) assess possible failure modes of an AI system. This will require a standardization of adversarial scoring and validation algorithms. The above-mentioned RobustML project [41], incidentally, can also be seen as an attempt to propose some standard for adversarial attacks and checks.

c) Expanding the range of adversarial examples: To date, much of the works on adversarial examples have focused on the invariance of the model to small perturbations, typically images. This has provided an excellent testing ground for developing approaches to adversarial assessment, robust learning, and validation, but alternative specifications that are directly relevant to the real world are needed.

We would like to highlight this point especially. The limited modification and widespread focus on image classification are some of the most serious limitations of current approaches.

At the same time, Deepmind notes that "manual" creation of specifications for AI systems will be difficult. Accordingly, systems are needed that can use partial “human” specifications and learn additional specifications based on evaluative feedback. Hence the focus on systems using reinforcement learning [44].

Other projects include the Fairness & Robustness in Machine Learning project of the Institute of Mathematics of the University of Toulouse [45]. The results presented relate more to the fairness and interpretability of the results.

ETH Zurich supports the Safe Artificial Intelligence project [46]. Judging by the publications presented, much attention is paid to the certification of the sustainability of machine learning systems. The project staff created Latticeflow, an industrial company [47], which positions itself as the world's first trusted AI platform that enables organizations to build and deploy robust AI models. Clients include Swiss Federal Railways (SBB) and Siemens, as well as government agencies such as the US Army and the German Federal Information Security Agency (BSI). The company does not provide a clear description of the product (however, this is typical for many software products related to safety), but the idea of the purpose and scheme of work can be understood from the existing description of the project to assess the sustainability of the railway sign recognition system [48]. The LatticeFlow system was used for testing when some standard transformations (rotations, changes in color and brightness, background, etc.)

were applied to existing data (images), after which the quality of recognition (classification) was evaluated with the new data.

Further, we can mention the project of the Alan Turing Institute on robust machine learning [49]. Safe and reliable AI is a priority area on the UK government's AI roadmap [50]. The works on the project are not published on the site, but there is at least one more program dedicated to adversarial machine learning [51]. This project also deals with the classification of images, belongs to the direction of Protection and Security. The latest works of the group members, according to Google Scholar, are devoted to the shift in distribution [52] (joint work with Yandex employees). The base for these projects is a research group in Oxford - Oxford Applied and Theoretical Machine Learning Group [53]. Her publications on the topic of adversarial learning are separated into a separate group [54].

The Allen Institute for AI (Paul Allen, Microsoft co-founder) [55] also supports adversarial research. An example is seminars at the University of Washington [56].

At the University of California, Berkeley, there is a group that is developing verifiable AI [57]. The research group behind this direction [58] both deals with formal methods of verification, as well as the development of appropriate applications (SMT solvers). Given the importance of formal verification methods for machine learning systems, this is one of the most interesting projects.

Industrial R&D centers include, for example, the Bosch Center for Artificial Intelligence [59]. Adversarial attacks are being considered by a group working on explainable deep learning models [60].

Ford has its own Core-Artificial Intelligence / Machine Learning (AI / ML) group. An example is a job ad for a robust machine learning job [61].

Yandex announced a competition to find solutions to efficiently work with a shift in distribution [62]. The objective is to raise awareness of shifts in the distribution of real data. The aim of the participants will be to develop models that are robust to a shift in the distribution and to identify such a shift using uncertainty measures in their predictions. Participants can take part in three separate tracks for which Yandex has provided datasets in three areas: weather forecast, machine translation, and vehicle forecast.

Startup Adversa [63] offers adversarial testing of AI systems. This is in line with what IT companies do, for example for third-party web services - penetration testing [64]. The company has published a rather interesting report on attacks [65].

Naturally, projects open up in the field of understanding (interpreting results) of machine learning systems. Or even more broadly, like Robust AI [66], an industrial-grade cognitive platform that empowers robots to reason with common sense. Here, research concerns not only (and not so much) deep learning, but also target symbolic AI.

From the point of view of the business model, this option opens up new forms of submission (presentation) of advertising content, which can be dynamic (support devices can, of course, change their mailings) and localized (available only in a limited area).

IV. Conclusion

As can be seen from the above review, they explicitly announced work in the field of verification of machine learning systems, that is, formal proof of their work with given data, which is the only “real” confirmation of robustness (real robustness in comparison with other methods) in Google Deepmind and the Berkeley University. At the same time, it should be noted that there are still few publications on the announced projects. This may be due to the fact that there are no significant results yet, or to restrictions on publication.

Perhaps for subsequent versions of the review, it is necessary to add an overview of patent applications, which may precede the publication of the results.

Image classification still dominates in the announced robust machine learning projects.

At the same time, it should be noted that the presence of adversarial examples is a fundamental characteristic of the current architecture of machine learning systems, when the data is divided, and training data, generally speaking, can differ arbitrarily from the general population. Accordingly, the need to work with a shift in the distribution, the need to build generalizations on the fly, using a small number of examples, should become part of the new model (s) of deep learning systems, as discussed in a recent article by Yoshua Bengio, Yann Lecun and Geoffrey Hinton (those who practically stood at the origins of the current models) [67].

Acknowledgment

We are grateful to the staff of the Department of Information Security of the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University for valuable discussions of this work.

References

[1] Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." IEEE Reviews in Biomedical Engineering 14 (2020): 156-180.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199 , 2013

[3] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, “Poison frogs! targeted clean-label poisoning attacks on neural networks,” in Advances in Neural Information Processing Systems, 2018, pp. 6103–6113.

[4] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” IEEE transactions on neural networks and learning systems, 2019.

[5] A Complete List of All (arXiv) Adversarial Example Papers https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. Retrieved: Aug, 2021

[6] Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: Aug, 2021

[7] Koh, Pang Wei, et al. "Wilds: A benchmark of in-the-wild distribution shifts." *International Conference on Machine Learning*. PMLR, 2021.

[8] Nair, Nimisha G., Pallavi Satpathy, and Jabez Christopher. "Covariate shift: A review and analysis on classifiers." 2019 Global Conference for Advancement in Technology (GCAT). IEEE, 2019.

[9] Major ML datasets have tens of thousands of errors https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors Retrieved: Aug, 2021
[10] NeuIPS https://www.llnl.gov/news/neurips-papers-aim-improve-understanding-and-robustness-machine-learning-algorithms
[11] Pei, Kexin, et al. "Towards practical verification of machine learning: The case of computer vision systems." arXiv preprint arXiv:1712.01785 (2017).
[12] Katz, Guy, et al. "The marabou framework for verification and analysis of deep neural networks." *International Conference on Computer Aided Verification*. Springer, Cham, 2019.
[13] Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." IEEE Design & Test 37.2 (2020): 30-57.
[14] The Pentagon Is Bolstering Its AI Systems—by Hacking Itself https://www.wired.com/story/pentagon-bolstering-ai-systems-hacking-itself Retrieved: Aug, 2021
[15] Poison in the Well Securing the Shared Resources of Machine Learning https://cset.georgetown.edu/publication/poison-in-the-well/ Retrieved: Aug, 2021
[16] Hacking AI A PRIMER FOR POLICYMAKERS ON MACHINE LEARNING CYBERSECURITY https://cset.georgetown.edu/wp-content/uploads/CSET-Hacking-AI.pdf Retrieved: Aug, 2021
[17] Guaranteeing AI Robustness Against Deception https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception Retrieved: Aug, 2021
[18] DARPA is pouring millions into a new AI defense program. Here are the companies leading the charge https://www.protocol.com/intel-darpa-adversarial-ai-project Retrieved: Aug, 2021
[19] UT Austin Selected as Home of National AI Institute Focused on Machine Learning https://news.utexas.edu/2020/08/26/ut-austin-selected-as-home-of-national-ai-institute-focused-on-machine-learning/ Retrieved: Aug, 2021
[20] UT Austin Launches Institute to Harness the Data Revolution https://ml.utexas.edu/news/611 Retrieved: Aug, 2021
[21] National Security Education Center https://www.lanl.gov/projects/national-security-education-center/ Retrieved: Aug, 2021
[22] 2021 Project Descriptions Creates next-generation leaders in Machine Learning for Scientific Applications https://www.lanl.gov/projects/national-security-education-center/information-science-technology/summer-schools/applied-machine-learning/project-descriptions-2019.php Retrieved: Aug, 2021
[23] Assured Machine Learning: Robustness, Fairness, and Privacy https://computing.llnl.gov/casc/ml/robust Retrieved: Aug, 2021
[24] Explainable Artificial Intelligence https://computing.llnl.gov/casc/ml/ai Retrieved: Aug, 2021
[25] Advancing Machine Learning for Mission-Critical Applications https://computing.llnl.gov/sites/default/files/COMP_ROADSHOW_ML_CASC-final.pdf Retrieved: Aug, 2021
[26] Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." Information Fusion (2021).
[27] Intelligence Advanced Research Projects Activity (IARPA) https://www.iarpa.gov/ Retrieved: Aug, 2021
[28] Trojans in Artificial Intelligence https://www.iarpa.gov/index.php/research-programs/trojai Retrieved: Aug, 2021
[29] Trojans in Artificial Intelligence bibliography https://scholar.google.com/scholar?hl=en&as_sdt=0%2C47&q=W911NF20C0034+OR+W911NF20C0038+OR+W911NF20C0045+OR+W911NF20C0035+OR+IARPA-20001-D2020-2007180011 Retrieved: Aug, 2021
[30] ELLIS Programs launched https://ellis.eu/news/ellis-programs-launched Retrieved: Aug, 2021
[31] ELLIS programs https://ellis.eu/programs Retrieved: Aug, 2021
[32] Robust Machine Learning https://ellis.eu/programs/robust-machine-learning Retrieved: Aug, 2021
[33] Semantic, Symbolic and Interpretable Machine Learning https://ellis.eu/programs/semantic-symbolic-and-interpretable-machine-learning Retrieved: Aug, 2021
[34] Oomen, Thomas L. "Why the EU lacks behind China in AI development–Analysis and solutions to enhance EU's AI strategy." rue 33.1: 7543.
[35] MIT Reliable and Robust Machine Learning https://www.csail.mit.edu/research/reliable-and-robust-machine-learning Retrieved: Aug, 2021
[36] Alexander Madry http://people.csail.mit.edu/madry/ Retrieved: Aug, 2021
[37] Center for Deployable Machine Learning (CDML) https://www.csail.mit.edu/research/center-deployable-machine-learning-cdml Retrieved: Aug, 2021
[38] Madry Lab http://madry-lab.ml/ Retrieved: Aug, 2021
[39] Robustness package https://github.com/MadryLab/robustness Retrieved: Aug, 2021
[40] Adversarial ML tutorial https://adversarial-ml-tutorial.org/ Retrieved: Aug, 2021
[41] RobustML https://www.robust-ml.org/ Retrieved: Aug, 2021
[42] Andriushchenko, Maksym, and Matthias Hein. "Provably robust boosted decision stumps and trees against adversarial attacks." arXiv preprint arXiv:1906.03526 (2019).
[43] Identifying and eliminating bugs in learned predictive models https://deepmind.com/blog/article/robust-and-verified-ai Retrieved: Aug, 2021
[44] Nandy, Abhishek, and Manisha Biswas. "Google's DeepMind and the Future of Reinforcement Learning." Reinforcement Learning. Apress, Berkeley, CA, 2018. 155-163.
[45] Fairness & Robustness in Machine Learning https://perso.math.univ-toulouse.fr/loubes/fairness-robustness-in-machine-learning/ Retrieved: Aug, 2021
[46] Safe Artificial Intelligence http://safeai.ethz.ch/ Retrieved: Aug, 2021
[47] Latticeflow https://latticeflow.ai/ Retrieved: Aug, 2021
[48] Reliability Assessment of Traffic Sign Classifiers https://latticeflow.ai/wp-content/uploads/2021/01/Reliability_assessment_of_traffic_sign_classifiers_short.pdf Retrieved: Aug, 2021
[49] The Alan Turing Institute Robust machine learning https://www.turing.ac.uk/research/interest-groups/robust-machine-learning Retrieved: Aug, 2021
[50] AI roadmap https://www.gov.uk/government/publications/ai-roadmap Retrieved: Aug, 2021
[51] The Alan Turing Institute Adversarial machine learning https://www.turing.ac.uk/research/research-projects/adversarial-machine-learning Retrieved: Aug, 2021
[52] Malinin, Andrey, et al. "Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks." arXiv preprint arXiv:2107.07455 (2021).
[53] Oxford Applied and Theoretical Machine Learning Group https://oatml.cs.ox.ac.uk/ Retrieved: Aug, 2021
[54] Adversarial and Interpretable ML — Publications https://oatml.cs.ox.ac.uk/tags/adversarial_interpretability.html#title Retrieved: Aug, 2021
[55] Allen Institute for AI https://allenai.org/ Retrieved: Aug, 2021
[56] AI2 Machine Learning Seminars https://www.cs.washington.edu/research/ml/seminars Retrieved: Aug, 2021
[57] Verified AI https://berkeleylearnverify.github.io/VerifiedAIWebsite/ Retrieved: Aug, 2021
[58] Sanjit A. Seshia research group https://people.eecs.berkeley.edu/~sseshia/ Retrieved: Aug, 2021
[59] Bosch AI https://www.bosch-ai.com/ Retrieved: Aug, 2021
[60] Rich and Explainable Deep Learning https://www.bosch-ai.com/research/research-fields/rich_and_explainable_deep_learning_perception/ Retrieved: Aug, 2021
[61] Research Engineer – Robust and Explainable AI Methods https://www.mendeley.com/careers/job/research-engineer-robust-and-explainable-ai-methods-690764 Retrieved: Aug, 2021
[62] Yandex Shift Challenge https://research.yandex.com/shifts Retrieved: Aug, 2021
[63] Adversa https://adversa.ai/ Retrieved: Aug, 2021
[64] De Jimenez, Rina Elizabeth Lopez. "Pentesting on web applications using ethical-hacking." 2016 IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI). IEEE, 2016.
[65] The Road to Secure and Trusted AI https://adversa.ai/report-secure-and-trusted-ai/ Retrieved: Aug, 2021
[66] Robust AI https://www.robust.ai/ Retrieved: Aug, 2021
[67] Bengio, Yoshua, Yann Lecun, and Geoffrey Hinton. "Deep learning for AI." Communications of the ACM 64.7 (2021): 58-65