On Audit and Certification of Machine Learning Systems

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

Manfred Sneps-Sneppe
Ventspils University of Applied Sciences
Venspils, Latvia
manfreds.sneps@gmail.com

***Abstract*—Obviously, machine learning applications are being used more and more in a wide variety of fields. The general rule today is that in the absence of analytical models, one always turns to machine learning. In itself, machine learning has become synonymous with artificial intelligence. The reverse is also true – artificial intelligence today is machine learning. Sometimes this definition is somewhat limited, and they only talk about artificial neural networks and deep learning in the context of artificial intelligence, but this does not change the essence of the matter. At the same time, it is also obvious that the spread of machine learning technologies leads to the need for their application in the so-called critical areas, where there are special requirements for confirming the operability and quality of software. These areas include, for example, avionics, nuclear power, autonomous vehicles, etc. Audit and, of course, certification are the procedures for evaluating machine learning models.**

I. Introduction

Machine learning systems are today the main examples of the use of Artificial Intelligence in a wide variety of areas. From a practical point of view, we can say that machine learning is synonymous with the concept of Artificial Intelligence. Yes, there is the concept of the so-called strong artificial intelligence (Strong AI, full AI, and AGI are also all synonyms), but it is still far from practical use. Accordingly, in practice, speaking of artificial intelligence systems, we should focus on the current architectures of machine learning systems, and on the available machine learning models and schemes for their implementation. Accordingly, the audit and certification of machine learning systems is now the same as the audit and certification of artificial intelligence systems.

Traditional software for mission-critical applications, such as avionics, goes through special certification procedures to ensure that it works. As shown in the paper, these ad hoc testing procedures cannot currently be directly transferred to machine learning models. However it is possible that a number of prescribed procedures (steps) must be followed to increase confidence in the results of the system. This, in fact, is an audit. The article discusses approaches to auditing and certification of machine learning models.

Considerations for certification of machine learning models were reported at the GRID-2023 conference [1]. Publications [2], [3] can be mentioned as previous works.

The study was supported by the Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence".

Machine learning models are data driven. Changing the data during the training phase, for example, leads to a change in the parameters of the model. Changing the input data (in relation to the data on which the model was trained) leads to a change in the results of the work. Such changes can be very significant and qualitative (for example, changing the classification of objects, etc.) or simply lead to a decrease in the accuracy of the system. Accordingly, based on this, the so-called adversarial attacks on machine learning models arise - conscious data modifications at different stages of the pipeline, which are designed to either interfere with the operation of the machine learning system, or, conversely, achieve the desired result for the attacker.

Google (Deepmind), in an overview post by its Robust and Verified Deep Learning group, notes that “machine learning systems are not trustworthy by default. Even systems that outperform humans in a certain area may fail to solve simple problems if differences are introduced into the input data” [4].

The Madry-lab (MIT) presentation introduced the three precepts of Secure/Safe ML [5]:

I. Thou shall not train data you don’t trust (because of data poisoning)
II. Thou shall not let anyone use your model (or observe its outputs) unless you completely trust them (because of model stealing and black box attacks)
III. Thou shall not fully trust the predictions of your model (because of adversarial examples)

Naturally, such adversarial examples are of particular importance for critical applications (avionics, automatic driving, nuclear power engineering, etc.). The consequences of errors here are always serious, and for such systems, there may be interested persons in such attacks.

A. On adversarial attacks

NIST, according to the latest recommendations [6], distinguishes three basic types of attacks against machine learning systems: poisoning [7], evasion [8], and attacks on intellectual property [9]. The latter are a special survey of models in order to extract non-public information and do not affect the results of the work (excluding conscious distortion of the output to counteract such attacks). The term poisoning is used to emphasize the long-term nature of the impact on models and includes data poisoning (special data modifications at the training stage) and model poisoning (direct modification of finished models [10]). Such attacks require access to training data (or loading poisoned data) or loading modified (poisoned)

models. In a first approximation, we can say that the requirements for protection against such attacks are similar to the usual requirements of cybersecurity (digital hygiene), with the prohibition of downloading anything from unknown sources (at least for critical applications this should definitely be excluded). What remains are evasion attacks, which consist in modifying (in the digital or physical domain) the input data. In the classical form, at the time of its appearance, these were the minimal modifications of the input data that caused the system to malfunction. Such an attack is depicted in Fig. 1, when noise is calculated using the gradient of the loss function, the addition of which to the original image changes the classification.

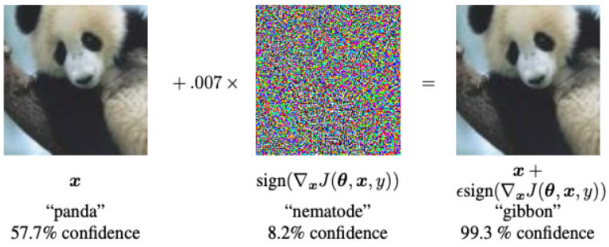

Fig. 1. On evasion attack [11].

The problem is that adversarial attacks are not necessary to achieve the indicated effect with a change in the classification. The trained model can be such that small changes in the data cause large changes in the result. And here we come to a much more important concept - these are robust machine learning models.

B. On robust machine learning

Classically, for machine learning systems, robustness is defined as the independence of the output of the system from small changes in the input data. The presence of such a dependency excludes, naturally, the use of such systems in applications where the results of work must be guaranteed. Robust machine learning models are a popular research topic, the basis for which was the need to use machine learning systems in critical systems [12].

Formally, robustness is defined approximately in the following form. Given an input x and a model of interest f, we want the model prediction to remain the same for all inputs x' in a neighborhood of x, where the neighborhood is defined by some distance function δ and some maximum distance Δ:

$$\forall x'.\ \delta(x, x') \leq \Delta \Rightarrow f(x) = f(x') \quad (1)$$

For example, the results of the classifier should not change with a small change in the data. The fundamental basis of robustness research is quite clear. Basically, any model is trained on some subset of data and then generalized to the entire population of data. This set, in the general case, is unknown at the time of training. And we turned to machine learning (artificial neural networks) precisely because the connections within the data are unknown to us. It is them that we want to restore (simulate) by training the neural network. This uncertainty suggests that the data during operation may differ from those on which the model was trained. Since the data has changed during operation, it may well turn out that the generalizations made during the training phase are no longer correct. If the data is changed in a special way, then this is called an attack on machine learning systems. It is around the formula (1) described above that all research in the field of stability is built. How to pick up minimally different data, which, nevertheless, is classified differently? Since in most cases we are talking about images, we are talking about changes imperceptible to the human eye, formally expressed in one of the L-metrics, which lead to a change in the classification. Or, in the opposite direction, check that the classification does not change with given small changes to input data.

Almost immediately, with such a formulation, the question arises - how does such a formulation correlate specifically with security? Suppose we have proved that in a small neighborhood of known data, the operation of the system remains stable. What happens outside of this area? How important is the "invisibility" of changes in general, if in critical applications (avionics, etc.) we are dealing with automatic systems, there is simply no person there, and the scope of changes, generally speaking, does not solve anything?

Everything looks so that small changes are chosen because it allows you to formally describe the ongoing processes and use previously known approaches. But this is not at all dictated by safety issues. It seems that in fact, at least for critical applications, robustness is interpreted (perceived) in a different form. Namely, the preservation of the performance indicators of the model, achieved at the stage of training, during its practical use [13].

There is a complete parallel with traditional software implementation. During the testing phase, we checked the performance of the system, and we expect this performance to continue during the operational phase. Note that for critical applications, the software is also subject to certification. The meaning of this certification is precisely in comprehensive testing (proof of correct operation). According to the same principle, stability is perceived. During the training phase, we have achieved certain selected performance indicators (accuracy, ROC, etc.) and expect the same parameters to be maintained when testing (exploiting) the model. For critical applications, the performance of the trained model below some certain level will simply be a stopping factor in the transition to production. That is, robustness becomes synonymous with performance. This is not the preservation of indicators for small perturbations of the training data, but the preservation of the indicators achieved at the stage of training already on the entire general population. And this is not at all what is usually studied in works on the stability of machine learning systems.

The same fact is noted in [14]. Robustness is a term that practitioners often use, but it usually refers generically to the correctness or validity of a model's predictions, rather than the formal notion of robustness (1) studied in the academic literature.

C. On robustness vs. safety

What is wrong, and what then is the point of robustness working at all? Confidence in the legitimacy of such questions was strengthened after reading the work [15], where Christian Kästner from Carnegie Mellon wrote exactly the same thing.

Note that the formula (1) does not say anything about the correct operation of the system (for example, about the

classification results). That is, there may well be a robust system that produces incorrect results. And these incorrect results remain so for small perturbations of the initial data. Hence, robustness by itself cannot be indicative of software safety. Safety (security) is a property of a system that includes a machine learning model. With regard to machine learning systems, security is used as a synonym for trust in the results of work [16]. Robustness from this practical point of view is inseparable from the explanation of how the model works. Indeed, by definition of the term "black box", we cannot guarantee any properties and characteristics for it, due to the fact that they are unknown and unpredictable.

The general state of affairs with discriminant machine learning models can be described simply. We can obviously get important results with machine learning (and this is the reason for the general interest), but, in general, these results cannot be guaranteed. Naturally, this is primarily a problem for critical systems. Software, for example in avionics, is certified to validate performance guarantees over the full range of possible inputs. The machine learning model at the stage of application (inference), in the same avionics, is also nothing but some software. And for such a program, certification is also required. The absence of such for artificial intelligence systems will lead to the division of special software into varieties - certified (verified) programs and uncertified ones.

Note that generative models (if we talk about LLM, which are now of the greatest interest) in this part do not differ from discriminant ones. Data poisoning attacks also exist for LLM [17]. The lack of verification is also noted in the OWASP list - "Over-reliance on LLM-generated content without human control can lead to detrimental consequences" [18]. Although in general, the main risks for large language models are currently seen in terms of access to accumulated information [40]. Therefore, in this article, we focus on discriminant models and classification systems, which are just typical for critical applications.

D. On contribution

We summarize our main contributions as follows. Based on the absolute need for certification of artificial intelligence (machine learning) systems, based on the completed systematization of knowledge (SOK), the article provides answers to the following questions:

1) Do the adopted and planned regulations relate to the certification of machine learning systems? (the answer is no);

2) Aare there enough robustness checks for secure systems? (no);

3) Is it currently possible to certify machine learning systems according to the same scheme as software in critical systems is certified? (no);

4) Certification of the robustness of machine learning models has very little to do with certification of software implementations of machine learning models (yes);

5) What is a necessary and feasible step towards certification of machine learning systems? (audit);

6) What could be the basis for an auditing standard?

The remainder of the article is structured as follows. In section II, we briefly dwell on the legal regulation in the field of AI, which, in fact, there are also regulations for machine learning systems. In Sections III and IV, we focus on audit and certification procedures, respectively. Section V is devoted to the basis (reasons) for certification. And section VI contains the conclusion.

II. ON LEGAL REGULATIONS FOR AI SYSTEMS

Ensuring performance assurance is naturally part of the various regulations for AI (ML) systems. As the MIT Technology Review notes in their collection The Algorithm "Suddenly, everyone wants to talk about how to regulate AI"

What's more, unusual in the industry, top executives are speaking out in favor of AI regulation. Executives from OpenAI, Microsoft, and Google have spoken out publicly in favor of regulation and held meetings with world leaders. And national governments are proposing new restrictions on generative AI. OpenAI CEO (Author of ChatGPT) Sam Altman went on a world tour to show support for new laws, including the upcoming European Union AI Act. OpenAI executives called on a global regulator to control superintelligent machines and testified in favor of AI regulation in front of the US Congress. The OpenAI company allocates grants for the development of AI control environments [19]. Microsoft President Brad Smith echoed OpenAI's calls for the US AI regulatory agency. Separately, Google CEO Sundar Pichai has agreed to work with European lawmakers to develop an "AI pact," a set of voluntary rules that developers must follow before EU rules come into effect. Undoubtedly, this process has accelerated precisely because of the success of large language models. At the annual meeting in Japan in 2023, the G7, an informal bloc of industrialized democratic governments, announced the Hiroshima process. It is an intergovernmental task force to investigate the risks of generative AI. Members of the G7 have pledged to develop mutually compatible laws that will allow AI to be regulated in line with democratic values. These include fairness, accountability, transparency, security, data privacy, abuse protection, and respect for human rights.

US President Joe Biden has published a strategic plan for the development of AI. The initiative calls on US regulators to develop publicly available datasets, benchmarks, and standards for training, measuring, and evaluating AI systems. France's data privacy regulator has announced a regulatory framework for generative AI. To date, China has already directly regulated generative AI. In March, EU officials rewrote the Union's law on artificial intelligence to classify generative artificial intelligence models as "high-risk", making them subject to bureaucratic oversight and regular reviews [20].

In the United States, the Algorithmic Accountability Act was introduced in Congress and the Senate in 2022. The bill would require companies to conduct algorithmic impact and risk assessments to address perceived harm from automated decision-making systems, such as those that deny people their mortgage applications.

The American Data Privacy Protection Act is an attempt to regulate the collection and processing of data by companies.

The debate around the risks of generative AI may give it added urgency. The law will prohibit companies engaged in generative artificial intelligence from collecting, processing, or transmitting data in a discriminatory manner. It will also give users more control over how companies use their data. For example, companies may be required to allow external experts to test their technologies before they are released, and to provide users and the government with more information about their AI systems.

The current debate among US lawmakers suggests that another agency to regulate AI is likely to emerge. It is also possible, according to the publication Algorithm, the emergence of a new controller specifically for AI tasks [21].

A new regulatory body set up by the European Union, the European Center for Algorithmic Transparency (ECAT), will study the algorithms that identify, classify, and rank information on social networking sites and search engines. ECAT is empowered to determine whether algorithms (AI and others) comply with the European Union's Digital Services Act, which aims to block online hate speech. The law should block certain content. The agency has three main tasks:

- Investigation. This is an evaluation of the functioning of "black box" algorithms. Includes analysis of reports and audits conducted by companies that are required by law to report to regulators. It will establish procedures for independent researchers and regulators to gain access to data related to algorithms.
- Study. This is an analysis of the capabilities of recommendation algorithms for the dissemination of illegal content, violation of human rights, damage to democracy or harm to the health of users, risk assessment and measures to reduce them, and increasing the transparency of algorithms (that is, explaining their work).
- Creation of a clearinghouse of information and best practices between researchers in academia, industry, and the public service.

And, of course, the adopted Law on Artificial Intelligence [20] should be mentioned here. The European Parliament has determined that secure AI developed (as well as used) in Europe must fully comply with EU rights and values, including human rights, security, privacy, transparency, non-discrimination, and social and environmental well-being.

AI systems with an unacceptable level of risk to human security will be banned. This category includes, for example, those that are used to classify people on the basis of their social behavior or some other personal characteristics (the so-called social assessment).

The law expands the restrictive list of prohibitions on the intrusive and discriminatory use of AI. This includes, for example, face recognition (in the text - real-time remote biometric identification), as well as biometric categorization (that is, categorization by gender, race, etc.). Also noted are predictive police systems, and emotion recognition systems in law enforcement, workplaces, and educational institutions. The law prohibits the inappropriate extraction of facial images from the Internet or video recordings from surveillance cameras to create facial recognition databases. For generative AI, the use of any copyrighted material in the training set of large language models, such as OpenAI GPT-4, is introduced.

High-risk applications include AI systems that cause significant harm to human health, safety, fundamental rights, or the environment, and systems used to influence voters and election outcomes. Also, note that the recommender systems used by social media platforms are classified as high-risk applications.

All of these acts define the requirements for the finished product. They do not define the practical steps to achieve the required performance, nor do they define the metrics by which those performances should be measured. It is the concepts of audit and certification that are already directly related to the practical area. Classically: an audit is an inspection (verification) process, that lists the requirements for checking AI systems (machine learning systems), and certification is already a confirmation (guarantee) of data (work results).

III. On audit of machine learning systems

Auditing machine learning systems is a new and fairly rapidly developing area. The reasons are the above problems with guaranteeing the results of the work. Among the 9 technologies that will change every industry, the Game Changers report names AI audit in the first place [24]. In the most recent areas, Bloomberg reports that China's Cyberspace Administration has announced draft guidelines that would require a security review of generative AI services before they are allowed to operate. The proposed rules say that AI operators must ensure content is accurate, respect intellectual property, do not jeopardize security, and do not discriminate. In addition, AI-generated content should be clearly tagged. The move is part of China's growing efforts to manage the rapid spread of generative AI since the debut of ChatGPT OpenAI last year.

In fact, an audit for machine learning systems is a set of best practices about what and how to check for finished systems. Proactively, it should also be practicing on how to develop secure systems. Today, we can say that the developers' understanding of the fact that such practices are needed clearly prevails over the understanding of what exactly needs to be done [23]. For example, here are the first 10 practices from this work:

Risk assessment before system deployment

Hazard Opportunity Ratings

Audit of third-party models

Security testing (red team)

Security restrictions

Model Verification Techniques

Security Incident Response Plan

Pre-training risk assessment

Monitoring systems and their use

Model evaluations after deployment

In fact, these are quite general points of the work plan. For most of them, there are no comprehensive (closing) solutions.

An audit for a machine learning (AI) system is an assessment of its algorithms, models, data, and design processes. This evaluation of AI applications by internal and external auditors helps to justify the robustness of the AI system, demonstrate the accountability of designers, and increase the validity of model predictions. AI audit covers [25]:

- Evaluation of models, algorithms, and data flows
- Analysis of operations, results, and detected anomalies
- Technical aspects of AI systems for evaluating the accuracy of results
- Ethical aspects of AI systems for fairness, legality, and privacy

This is in line with conventional definitions that an audit is a tool for questioning complex processes to determine whether they comply with company policy, industry standards, or regulations [33]. The IEEE Standard for Software Development defines an audit as “an independent assessment of the conformity of software products and processes with applicable codes, standards, guidelines, plans, specifications, and procedures” [34].

Below are sample sections of an AI audit and the issues covered [25]:

ML modeling: alternate ML approaches; reasons to justify chosen ML strategy; refining ML algorithms.

AI project scope definition: constraints and other implementation approaches.

Deployment and testing: methods used to deploy ML models; post deployment review; metric used to ensure accuracy of ML models.

Data management: data sources and data consistency; data imputation and data standardization.

Data monitoring: monitor model performance, drift, activities, and anomalies; compliance with law and regulatory standards; ethical and social responsibility.

Again, AI-audit is exactly the checklist for checking the availability of the necessary activities. For example, let's take the ML monitoring section. Here it is proposed to find answers to the following questions:

- Does the AI system have an appropriate monitoring process to track model performance, deviations, and model actions?
- What actions are taken in the execution of the machine learning pipeline to ensure that AI applications comply with laws and regulatory standards, meet organizational goals, and demonstrate ethical and social responsibility?

The issue of monitoring machine learning models is quite complex in itself. Depending on the type of possible data shift [26], monitoring conclusions will be different. And depending on the nature of the systems, solutions can also be different. For an application that runs 24x7, for example, a so-called concept shift is a disaster because such applications cannot be stopped to retrain.

At the same time, just as a checklist that defines the mandatory steps in development (operation), these are quite working recommendations. Compliance with the given positions is assessed “manually”. An example is the work of Stanford University [30] on the assessment of compliance with the draft European law on AI (Fig. 2). Compliance assessments were made manually by experts.

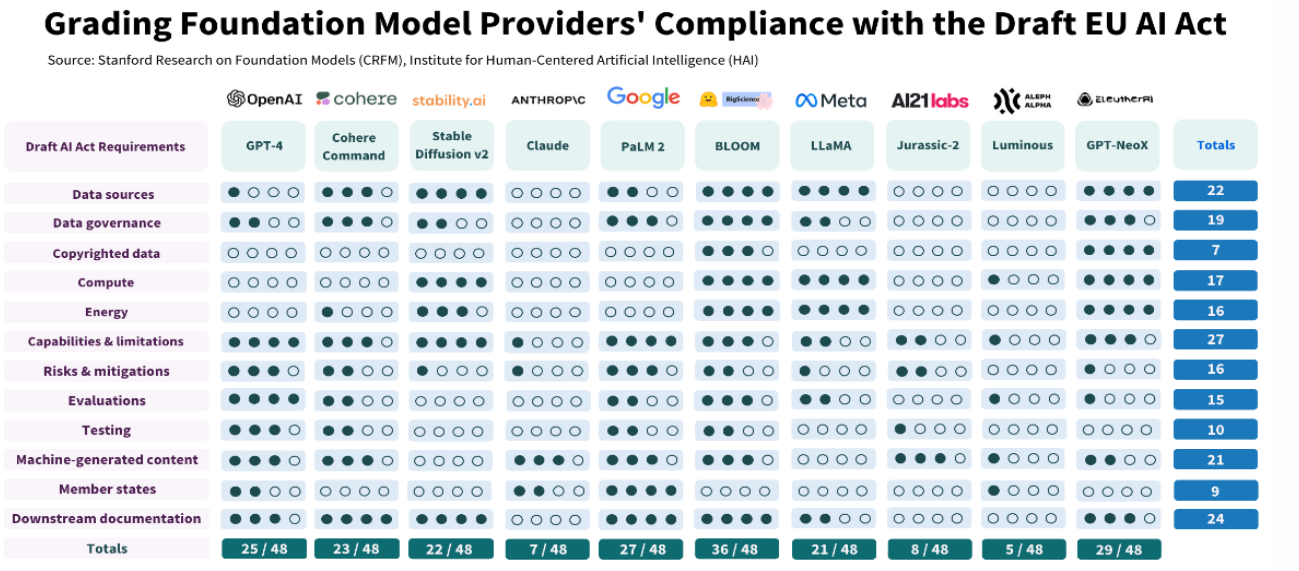

Fig. 2. Language models and the European AI law [30].

There are ready-made frameworks that help you navigate these tasks. Some of them are well-known frameworks used to describe IT assets. For example, COBIT [27]. There are frameworks focused directly on AI tasks, for example, the IIA Artificial Intelligence Auditing Framework [28] or Deloitte's Trustworthy AI Framework [29].

Speaking of possible standardization, the NIST AI RISC Management framework [31] and ISO/IEC 23894 [32] can be mentioned. Work on the creation of audit frameworks continues. Another such project is presented in the work of Google employees [35]. In Fig. 3 from this article, the gray color indicates a process, and the colored sections represent documents. Documents highlighted in orange are written by the auditors, while documents in blue are written by the development and product team. "Green products" are being developed jointly.

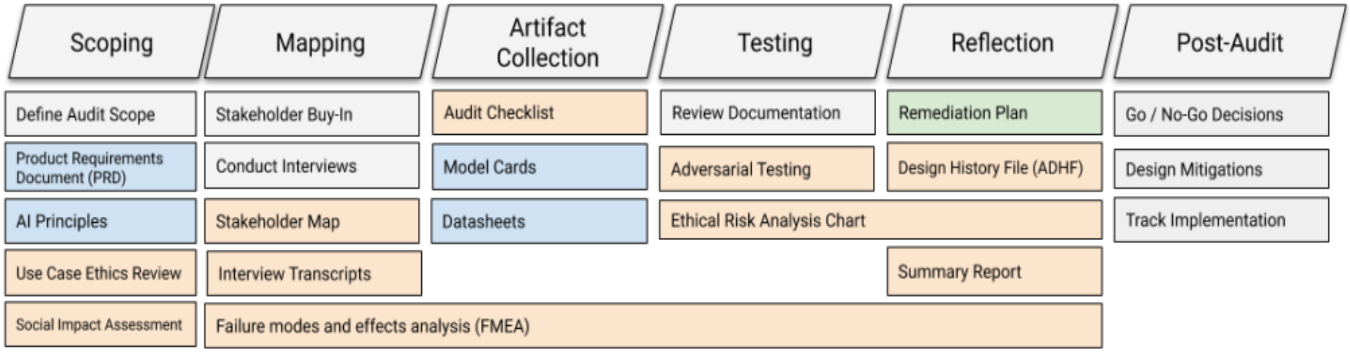

Fig. 3. Framework for auditing AI algorithms [35].

Another example also applies to Google (Deepmind) [36] is a work on a framework for evaluating the extreme risks of multipurpose AI systems [37]. The work is inspired by the practical implementation of ChatGPT and other large language models. Despite the already sufficient history of audit frameworks, this work of 2023 is positioned by the authors as the first. In fact, this is a set of statements about the need for responsible training of the model and the corresponding implementation. A universal framework for auditing AI systems is proposed in [38] by the German industrial company TUV. Its main content takes into account textual requirements such as "The model must be resistant to the PGD attack with a budget $\epsilon = 0.75$" and "ML model must use no more than 15% of background information for classification".

Research on the audit of AI systems (ML) is being conducted at the Fraunhofer organization [39]. The released Manifesto on Audited AI Systems [42] proposes an assessment matrix (Fig. 4) similar to that used in Stanford's paper [30] mentioned above.

Fig. 4. Estimates of auditability [42].

Gartner proposed AI TRiSM (Artificial Intelligence (AI) Trust, Risk, and Security Management) as a framework that provides AI management, reliability, fairness, efficiency, and privacy [41]. AI TRiSM focuses on:

- Trust in AI systems
- Risks of AI systems
- AI security management

In addition, Gartner defines 5 basic elements of AI TRiSM on which to build effective AI solutions:

- Explainability
- ModelOps - according to Gartner is the leadership and lifecycle management of artificial intelligence (AI) models and decision models, including machine learning, knowledge graphs, rules, optimization, linguistic and agent-based models. Key features include continuous integration, model development environments, testing, model versioning, and model storage.
- Data anomaly detection
- Countering adversarial attacks
- Data protection

The topic of audit, technically, should also include the so-called trust systems (platforms) for developing AI applications [16]. The very idea of trusted platforms in computer science is not new. The main point of trusted computing is to give hardware manufacturers control over what software works (does not work) on a system by refusing to run unsigned software. With trusted computing, the computer will always behave in the expected way, and that behavior will be enforced by the computer hardware and software. Ensuring this behavior is achieved by loading the hardware with a unique encryption key that is not available to the rest of the system and its owner. This concept is also necessary for machine learning systems in critical applications, since there are, for example, attacks that target machine learning frameworks. Changing, for example, the loss calculation function in a particular framework will affect all machine learning models on such a platform [6]. But for machine learning systems, this is only the smallest of the problems. The main problem of distrust comes precisely from the lack of trust in data processing. And trusted platforms are platforms whose tools allow you to increase confidence in machine learning models, platforms that allow you to analyze training data, resist adversarial attacks, determine data shifts during system operation, etc. Examples of such platforms are Datarobot [43] and IBM Trustworthy [44].

IV. On certification of machine learning systems

As noted earlier, certification is already a guarantee of the results of the system. Here it is necessary to dwell on the possible (existing) in fact different interpretations of the concept of certification for ML systems and software systems. For ML systems (models), certification is obtaining estimates of selected metrics (including probabilistic estimates). For programs, this is a guarantee of performance. Literally: “avionic systems should safely perform their intended function under all foreseeable operating and environmental conditions”. The machine software model in the inference phase is the program. And, accordingly, it must be certified, like any other program for critical applications. And probabilistic estimates, for example, do not work here at all.

How can machine learning systems be guaranteed to work?

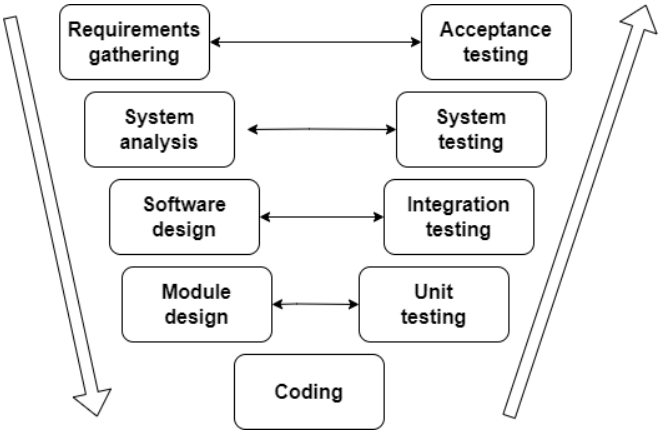

Fig. 5. V-model of the life cycle

Software Assurance (SwA) is a critical software development process that ensures that software products are reliable, safe, and secure. It includes many activities: requirements analysis, design analysis, code review, testing, and formal review. One of the most important components of software security is secure coding practices that follow industry standards and best practices.

There is a classical V-model of software development [45]. Two test directions (Fig. 5)

- Verification – are we building the product correctly?
- Validation - is the correct product built?

Each level has a corresponding set of tests. During verification, it is checked whether the product meets the requirements: it has all the functions for its intended use, as described in the planning phase after verification with its potential users, and these functions work as intended. This implies, firstly, the establishment of requirements, and then the creation of a system design specification based on them. Then the development moves in depth, refining the data of the previous step. During validation, it is checked whether the requirements describe what is really needed, whether they correctly take into account the goals of interested parties, and whether the received software corresponds to the application model.

For machine learning systems (neural networks), there are obviously components that can be tested in a similar way. For example, analysis of input data, monitoring of system operation, etc. But the key function (output) cannot be verified in this way (line by line). Daedalan and EASA (European Union Aviation Safety Agency) proposed the term Learning Assurance instead of Software Assurance [46] and the corresponding W-model (Fig. 6).

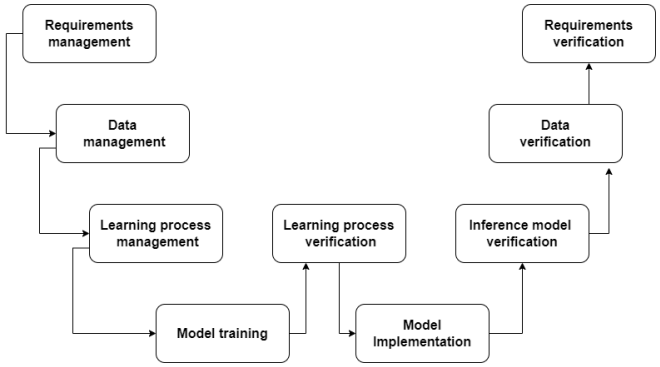

Fig. 6. W-model

This has been published as Concepts of Design Assurance for Neural Networks or CoDANN [47]. This concept may form the basis for future regulatory requirements. The EASA concept for certification of ML applications is outlined in their reports [54], [55], [56]. In our opinion, the element of this roadmap "EASA concept paper: First usable guidance for level 1 machine learning applications" is the best document to date, which is an audit of machine learning systems.

How is the W-model fundamentally different from the V-model? For machine learning systems, each step in Fig. 8 exists on its own. For example, it is assumed that we will examine the datasets and fix the "correct" option. Next, for a fixed dataset, we debug the model. After receiving satisfactory metrics, the model is fixed, and so on. That is, the problem is reduced to a sequence of deterministic steps (at each step, some deterministic result is obtained). It turns out that the question of a possible shift of data generally falls out of consideration. And how this process will work in the case of, for example, a shift in concepts [26] is not at all clear.

EASA has published a roadmap for its certification projects [48]. The latest version is dated May 2023. AI applications for aviation are divided into 3 levels:

Level 1 – assistance to human:

- Level 1A: Human augmentation
- Level 1B: Human cognitive assistance in decision-making and action selection

Level 2 – human-AI teaming

- Level 2A: Human and AI-based system cooperation
- Level 2B: Human and AI-based system colloboration

Level 3 – advanced automation

- Level 3A: The AI-based system performs decisions and actions that are overridable by the human
- Level 3B: The AI-based system performs non-overridable decisions and actions (e.g. to support safety upon loss of human oversight)

Certification of applications of the first level (human assistants) refers to 2025, and the last third level (non-cancellable actions) - to 2035-2050.

There are SAE (Society of Automotive Engineers) standards dedicated to artificial intelligence [49]. The G-34 Aviation AI Committee is responsible for creating and maintaining SAE technical reports (i.e. aerospace information reports, aerospace recommended practices, and aerospace standards) on implementation and certification aspects related to AI technologies, including any airborne systems for the safe operation of aerospace systems and aerospace vehicles. Working groups include all necessary committees:

- SG1 - Airborne & Ground Applications
- SG2 - ML Data Management & Validation
- SG3 - ML Design & Verification
- SG4 - ML Implementation & Verification
- SG5 - System & Safety Considerations for ML
- SG7 - Process Considerations (Planning, Config. Mgmt., Quality, Leveling, and Certification/Approval)

But there is only one publicly released paper from 2021, which is titled Artificial Intelligence in Aeronautical Systems: Statement of Concerns. Basically, the title of this document accurately describes the current state of the certification process.

In [50], it is noted that a quantitative assessment of the safety of AI in aviation is still under development. Literally: "The process of machine learning is by its nature very non-deterministic - it should be so, at least during the training phase. However, during the deployment phase, the inference engines that run, for example, convolutional neural networks and use this "learned information" can be customized to suit the requirements of certification authorities. Not everyone realizes this or deals with it, but at some point, inference engine determinism must be addressed by every system that wants to achieve a high level of security criticality." The requirement of determinism is understandable, but it conflicts with the main concept of machine learning - we believe that the generalizations developed at the training stage on the training data set will remain such for the entire general population. In the general case, without some external restrictions on this population (that is, on valid data), this cannot be guaranteed.

In [51], the authors, Airbus employees note that the certification of machine learning systems is a complex problem. This certification includes, at least, robustness certification, data validation, model provability, and model explainability. Data validation, for example, includes data fairness, data correctness, lineage, and competeness. Explainability, in turn, includes at least local and global explanation, feature importance, influence functions, saliency maps, etc.

In this scheme, data analysis, for example, is not only a static analysis of training datasets. Of course, they need to be analyzed, since, for example, backdoors in the model may be due to specially prepared data at the training stage. But data analysis is also necessary at the stage of model operation. As a

result of the complexity of such a process, in many works, it is simplified and reduced to robustness analysis [51].

The key role of robustness is twofold: on the one hand, and according to ED-12C/DO-178C, it is the degree to which software can continue to work correctly despite abnormal inputs and conditions. On the other hand, and more specifically for an ML application, EASA [52] defines that an ML system is robust when it produces the same output for inputs varying in the state-space region. Variations (disturbances) can be natural (for example, sensor noise, measurement bias, etc.), variations due to failures (for example, incorrect data from corrupted sensors), or deliberately inserted (for example, altered pixels in images) to deceive model predictions. When perturbed examples cheat the ML algorithm, we are talking about adversarial examples. This is usually defined as noise on the inputs that is imperceptible or does not exceed a threshold.

In [53], the authors dwell in sufficient detail on the fundamental incompatibility of the development process of ML applications and the provisions of DO-178. The main inconsistencies can be represented as follows:

- deterministic approach in certification of software systems against non-deterministic ML models
- code coverage (it is based on V-model - why is this line in the code?)
- data coverage. A standard approach in ML is point-wise robustness. Certification for ML models is a study of robustness in a limited range of modifications of correct data. E.g. 35.42% certified accuracy on MNIST under perturbation ϵ= 8/255. Here is the data set (correct images) and the limits of change for the pixel ($\pm$ 8)

V. On Technical Grounds for Certification

We can distinguish two approaches. Firstly, it is a formal verification of machine learning models. This is a working approach [57] with one big problem - scalability. Verification of models is reduced to logical formulas, or to systems of linear equations, and with an increase in the number of parameters (and now it is no longer millions of parameters), the task becomes unsolvable.

Another approach is to guarantee (certify) robustness [58]. Robustness testing approaches aim to assess the robustness of neural networks by providing a theoretically validated lower bound on robustness under certain perturbation constraints.

Outside of theoretical assessments, there is also the so-called adversarial learning, which aims to improve such a lower bound.

We can divide validation approaches into complete validation and incomplete validation. When the check method outputs "not checked" for a given x_0, if it is guaranteed that there is an adversarial instance of x around x_0, we call this a complete check, otherwise, it is an incomplete check.

We can also divide verification approaches into deterministic verification and probabilistic verification. When given inputs are not resistant to attack, deterministic testing guarantees the output "not tested", and probabilistic testing guarantees the output "not tested" with a certain probability (for example, 99.9%), where the randomness does not depend on the input data.

The final picture from [58] is perhaps the most complete description of the existing methods to date.

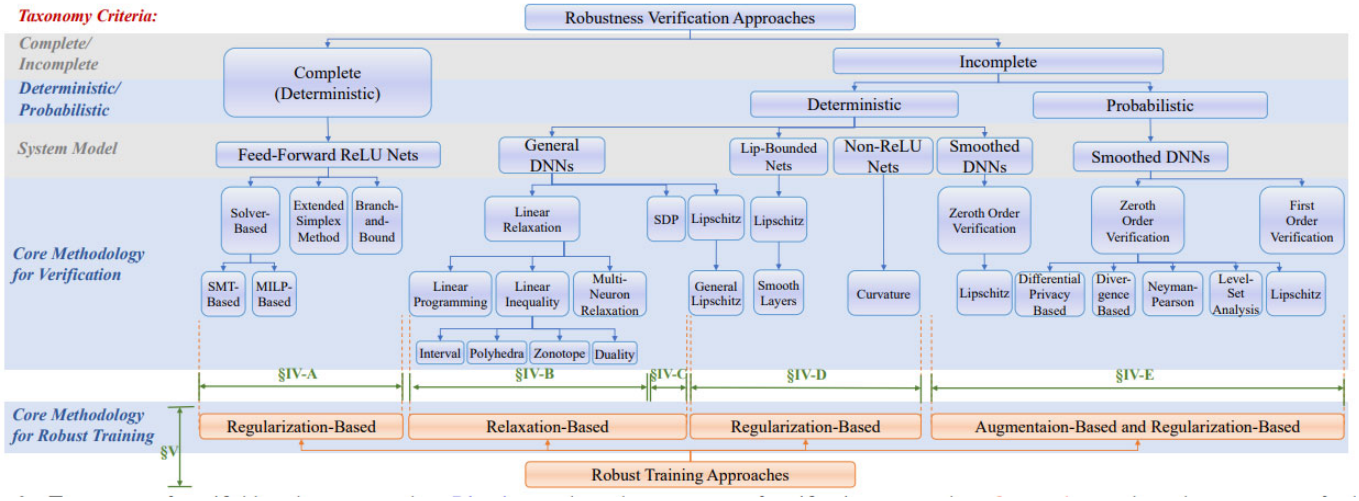

Fig. 7. Certified robustness [58]

But in this whole big tree, only the leftmost edge is suitable for software certification (complete and deterministic approaches). In other words, only a small part of the current work on the certification of machine learning models is relevant to what is currently understood as software certification.

In addition, the following must be kept in mind. Fig. 8 shows the fragment of results for certified accuracy in the L_{inf} metric (which is what arbitrary change testing is) from the paper [59]. We have highlighted the last column in the table above - this is exactly the data that is interesting to us (certified accuracy). As

you can see, these values are far from the typical values used in technical systems (so-called nines).

In other words, even based on the limited certification of the models themselves, the results do not provide a basis for using these models in systems that are certified to modern technical safety standards.

Dataset	Method	FLOPs	Test	Robust	Certified
MNIST ($\epsilon = 0.3$)	Group Sort (Anil et al., 2019)	2.9M	97.0	34.0	2.0
	COLT (Balunovic & Vechev, 2020)	4.9M	97.3	-	85.7
	IBP (Gowal et al., 2018)	114M	97.88	93.22	91.79
	CROWN-IBP (Zhang et al., 2020b)	114M	98.18	93.95	92.98
	ℓ_∞-dist Net	82.7M	98.54	94.71	92.64
	ℓ_∞-dist Net+MLP	85.3M	**98.56**	**95.28**	**93.09**
Fashion MNIST ($\epsilon = 0.1$)	CAP (Wong & Kolter, 2018)	0.41M	78.27	68.37	65.47
	IBP (Gowal et al., 2018)	114M	84.12	80.58	77.67
	CROWN-IBP (Zhang et al., 2020b)	114M	84.31	80.22	78.01
	ℓ_∞-dist Net	82.7M	**87.91**	79.64	77.48
	ℓ_∞-dist Net+MLP	85.3M	**87.91**	**80.89**	**79.23**
CIFAR-10 ($\epsilon = 8/255$)	PVT (Dvijotham et al., 2018a)	2.4M	48.64	32.72	26.67
	DiffAI (Mirman et al., 2019)	96.3M	40.2	-	23.2
	COLT (Balunovic & Vechev, 2020)	6.9M	51.7	-	27.5
	IBP (Gowal et al., 2018)	151M	50.99	31.27	29.19
	CROWN-IBP (Zhang et al., 2020b)	151M	45.98	34.58	33.06
	CROWN-IBP (loss fusion) (Xu et al., 2020a)	151M	46.29	35.69	33.38
	ℓ_∞-dist Net	121M	**56.80**	**37.46**	33.30
	ℓ_∞-dist Net+MLP	123M	50.80	37.06	**35.42**

Fig. 8. On certified accuracy [59].

VI. Conclusion

What do we have, as a result, today? And, accordingly, what can be done already now to guarantee the results of machine learning systems?

First, it should be noted that legal regulations only describe the final requirements for products and, accordingly, have nothing to do with either the process of achieving (satisfying) these requirements, or the procedure for conformity testing itself. The regulations in their current form define some recommendations, but they should be considered fairly obvious, if not trivial. For example, the need for explanations for artificial intelligence models.

Certification of machine learning systems, as it is understood for traditional software, is generally not possible today. A working deterministic approach is the formal verification of machine learning models, but it has scalability issues. Perhaps the solution to certify machine learning systems is to change existing standards. From a practical point of view, the certification of machine learning models is the certification of robustness, when metrics are guaranteed for a given budget (size) of training data modification. Another approach used in practice is to simulate possible problems with measuring devices (e.g. cameras). This can be called semantically driven change. It is also promising to study the problem of global evasion attacks.

The next stage is the audit of machine learning systems. From a practical point of view, an audit is, first of all, a checklist that lists the list of necessary actions (procedures) at different stages of the standard pipeline of machine learning models. The activities entail the creation of documents that describe the characteristics of the models being audited. These are practical and absolutely feasible procedures today that should be put into practice for all industrial machine learning models. In our opinion, the EASA concept paper [56] can be used specifically for auditing and serve as the basis for corporate (industry or even national) audit systems.

References

[1] GRID 2023 https://indico.jinr.ru/event/3505/ Retrieved: Aug, 2023

[2] Namiot, D., E. Ilyushin, and I. Chizov. "On errors and failures of machine learning projects." *AIP Conference Proceedings*. Vol. 2812. No. 1. AIP Publishing, 2023.

[3] Sneps-Sneppe, Manfred, and Dmitry Namiot. "On Artificial Intelligence: Software and Statistical Issues." *PROCEEDING OF THE 32ND CONFERENCE OF FRUCT ASSOCIATION*, pp.394-402, 2022

[4] Robust and Verified Deep Learning group https://deepmindsafetyresearch.medium.com/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda Retrieved: Aug, 2023

[5] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: Aug, 2023

[6] Namiot, Dmitry. "Schemes of attacks on machine learning models." *International Journal of Open Information Technologies* 11.5 (2023): 68-86. (in Russian)

[7] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." *International Journal of Open Information Technologies* 11.3 (2023): 58-68. (in Russian)

[8] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." *International Journal of Open Information Technologies* 10.10 (2022): 11-20. (in Russian)

[9] Song, Junzhe, and Dmitry Namiot. "A Survey of the Implementations of Model Inversion Attacks." *Distributed Computer and Communication Networks: 25th International Conference*, DCCN 2022, Moscow, Russia, September 26–29, 2022, Revised Selected Papers. Cham: Springer Nature Switzerland, 2023.

[10] Bidzhiev, Temirlan, and Dmitry Namiot. "Research of existing approaches to embedding malicious software in artificial neural networks." *International Journal of Open Information Technologies* 10.9 (2022): 21-31. (in Russian)

[11] Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. arXiv, 2014; arXiv:1412.6572.

[12] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." *International Journal of Open Information Technologies* 9.10 (2021): 35-46. (in Russian)

[13] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." *International Journal of Open Information Technologies* 10.9 (2022): 126-134. (in Russian)

[14] Borg, Markus, et al. "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry." arXiv preprint arXiv:1812.05389 (2018)

[15] Why Robustness is not Enough for Safety and Security in Machine Learning https://towardsdatascience.com/why-robustness-is-not-enoughfor-safety-and-security-in-machine-learning-1a35f6706601 Retrieved: Aug, 2023

[16] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." *International Journal of Open Information Technologies* 10.7 (2022): 119-127. (in Russian)

[17] Gu, Kang, et al. "Towards Sentence Level Inference Attack Against Pre-trained Language Models." *Proceedings on Privacy Enhancing Technologies* 3 (2023): 62-78.

[18] OWASP Top 10 List for Large Language Models version 0.1 https://owasp.org/www-project-top-10-for-large-language-model-applications/descriptions/

[19] Democratic inputs to AI https://openai.com/blog/democratic-inputs-to-ai Retrieved: Aug, 2023

[20] The AI Act https://artificialintelligenceact.eu/ Retrieved: Aug, 2023

[21] AI regulation https://www.technologyreview.com/2023/05/23/1073526/suddenly-everyone-wants-to-talk-about-how-to-regulate-ai/ Retrieved: Aug, 2023

[22] China Mandates Security Reviews for AI Services Like ChatGPT https://www.bloomberg.com/news/articles/2023-04-11/china-to-mandate-security-reviews-for-new-chatgpt-like-services

[23] Schuett, Jonas, et al. "Towards best practices in AGI safety and governance: A survey of expert opinion." arXiv preprint arXiv:2305.07153 (2023).

[24] Game Changers https://www.cbinsights.com/research/report/game-changing-technologies-2022/ Retrieved: Aug, 2023

[25] An In-Depth Guide To Help You Start Auditing Your AI Models https://censius.ai/blogs/ai-audit-guide Retrieved: Aug, 2023

[26] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." *International Journal of Open Information Technologies* 10.12 (2022): 84-93. (in Russian)

[27] van Wyk, Jana, and Riaan Rudman. "COBIT 5 compliance: best practices cognitive computing risk assessment and control checklist." Meditari Accountancy Research (2019).

[28] The IIA's Artificial Intelligence Auditing Framework https://www.theiia.org/en/content/articles/global-perspectives-and-insights/2017/the-iias-artificial-intelligence-auditing-framework-practical-applications-part-ii/ Retrieved: Aug, 2023

[29] REALIZE THE FULL POTENTIAL OF ARTIFICIAL INTELLIGENCE https://www.coso.org/Shared%20Documents/Realize-the-Full-Potential-of-Artificial-Intelligence.pdf Retrieved: Aug, 2023

[30] Do Foundation Model Providers Comply with the Draft EU AI Act? https://crfm.stanford.edu/2023/06/15/eu-ai-act.html Retrieved: Aug, 2023

[31] AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework Retrieved: Aug, 2023

[32] ISO/IEC 23894 – A new standard for risk management of AI https://aistandardshub.org/a-new-standard-for-ai-risk-management Retrieved: Aug, 2023

[33] Jie Liu. 2012. The enterprise risk management and the risk oriented internal audit. *Ibusiness* 4, 03 (2012), 287.

[34] IEEE. 2008. IEEE Standard for Software Reviews and Audits. *IEEE Std* 1028-2008 (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[35] Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.

[36] New research proposes a framework for evaluating general-purpose models against novel threats https://www.deepmind.com/blog/an-early-warning-system-for-novel-ai-risks Retrieved: Aug, 2023

[37] Shevlane, Toby, et al. "Model evaluation for extreme risks." arXiv preprint arXiv:2305.15324 (2023).

[38] Markert, Thora, Fabian Langer, and Vasilios Danos. "GAFAI: Proposal of a Generalized Audit Framework for AI." *INFORMATIK* 2022 (2022).

[39] Auditing and Certification of AI Systems https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html Retrieved: Aug, 2023

[40] Derner, Erik, and Kristina Batistič. "Beyond the Safeguards: Exploring the Security Risks of ChatGPT." arXiv preprint arXiv:2305.08005 (2023).

[41] AI TRiSM https://www.gartner.com/en/information-technology/glossary/ai-trism Retrieved: Aug, 2023

[42] Towards Auditable AI Systems Retrieved: Aug, 2023 https://www.hhi.fraunhofer.de/fileadmin/Departments/AI/TechnologiesAndSolutions/AuditingAndCertificationOfAiSystems/2022-05-23-whitepaper-tuev-bsi-hhi-towards-auditable-ai-systems.pdf Retrieved: Aug, 2023

[43] Datarobot https://www.datarobot.com/platform/trusted-ai/ Retrieved: Aug, 2023

[44] IBM Trustworthy https://research.ibm.com/topics/trustworthy-ai Retrieved: Aug, 2023

[45] Ruparelia, Nayan B. "Software development lifecycle models." *ACM SIGSOFT Software Engineering Notes* 35.3 (2010): 8-13.

[46] Explaining W-shaped Learning Assurance https://daedalean.ai/tpost/pxl6ih0yc1-explaining-w-shaped-learning-assurance Retrieved: Aug, 2023

[47] Force, DA EASA AI Task, and A. G. Daedalean. "Concepts of Design Assurance for Neural Networks (CoDANN)." Concepts of Design Assurance for Neural Networks (CoDANN). EASA, Daedalean (2020).

[48] EASA roadmap https://www.easa.europa.eu/en/domains/research-innovation/ai Retrieved: Aug, 2023

[49] G-34 Artificial Intelligence in Aviation https://standardsworks.sae.org/standards-committees/g-34-artificial-intelligence-aviation Retrieved: Aug, 2023

[50] DO-178 continues to adapt to emerging digital technologies https://militaryembedded.com/avionics/safety-certification/do-178-continues-to-adapt-to-emerging-digital-technologies Retrieved: Aug, 2023

[51] Vidot, Guillaume, et al. "Certification of embedded systems based on Machine Learning: A survey." arXiv preprint arXiv:2106.07221 (2021).

[52] EASA Artificial Intelligence Roadmap 1.0. https://www.easa.europa.eu/sites/default/files/dfu/EASA-AIRoadmap-v1.0.pdf Retrieved: Aug, 2023

[53] Dmitriev, Konstantin, Johann Schumann, and Florian Holzapfel. "Towards Design Assurance Level C for Machine-Learning Airborne Applications." *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*. IEEE, 2022.

[54] "Concepts of design assurance for neural networks (CoDANN)," European Aviation Safety Agency, Tech. Rep., 2020.

[55] "Report. concepts of design assurance for neural networks (CoDANN) II," European Aviation Safety Agency, Tech. Rep., 2021.

[56] "EASA concept paper: First usable guidance for level 1 machine learning applications," European Aviation Safety Agency, Tech. Rep., 2021.

[57] Stroeva, Ekaterina, and Aleksey Tonkikh. "Methods for Formal Verification of Artificial Neural Networks: A Review of Existing Approaches." *International Journal of Open Information Technologies* 10.10 (2022): 21-29. (in Russian)

[58] Li, Linyi, Tao Xie, and Bo Li. "Sok: Certified robustness for deep neural networks." *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023.

[59] Zhang, Bohang, et al. "Towards certifying robustness using neural networks with l-dist neurons." arXiv preprint arXiv:2102.05363 (2021).