social sciences

MDPI

Article

Enabling Artificial Intelligence Adoption through Assurance

Laura Freeman *, Abdul Rahman and Feras A. Batarseh

Virginia Polytechnic Institute, State University (Virginia Tech), 900 N. Glebe Road, Arlington, VA 22203, USA; abdul@vt.edu (A.R.); batarseh@vt.edu (F.A.B.)
* Correspondence: laura.freeman@vt.edu

Abstract: The wide scale adoption of Artificial Intelligence (AI) will require that AI engineers and developers can provide assurances to the user base that an algorithm will perform as intended and without failure. Assurance is the safety valve for reliable, dependable, explainable, and fair intelligent systems. AI assurance provides the necessary tools to enable AI adoption into applications, software, hardware, and complex systems. AI assurance involves quantifying capabilities and associating risks across deployments including: data quality to include inherent biases, algorithm performance, statistical errors, and algorithm trustworthiness and security. Data, algorithmic, and context/domain-specific factors may change over time and impact the ability of AI systems in delivering accurate outcomes. In this paper, we discuss the importance and different angles of AI assurance, and present a general framework that addresses its challenges.

Keywords: AI assurance; data quality; operating envelopes; validation and verification; XAI; AI trustworthiness; data democracy

check for updates

Citation: Freeman, Laura, Abdul Rahman, and Feras A. Batarseh. 2021. Enabling Artificial Intelligence Adoption through Assurance. *Social Sciences* 10: 322. https://doi.org/10.3390/socsci10090322

Academic Editor: Nigel Parton

Received: 29 June 2021
Accepted: 19 August 2021
Published: 25 August 2021

1. Introduction

At the famous 1956 Dartmouth workshop, Artificial Intelligence (AI) was defined as the "*aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*." Since then, there have been multiple formal and informal definitions of AI. However, few can argue against the notion that AI has been rekindled due to the *Big Data* revolution where commodity parallel processing can be elastically scaled in the cloud. Accordingly, the latest reincarnation of AI was delivered through the availability of extensive compute, the ability to store massive amounts of data, and algorithms/models to support predictions, classifications, and correlations. This new wave of AI, characterized by DARPA as statistical learning Launchbury (2017), includes analysis tools based in learning from data via deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks (RNN), among others. This new wave of techniques has proven very effective at meeting and exceeding human level performance in specific tasks, for example, Amodei et al. (2016) show how deep learning (DL) techniques can improve accuracy across diverse languages, He et al. (2015) show deep learning can surpassed human-level performance in image classification, and Silver et al. (2016) show how algorithms are mastering playing games. However, these improvements are not without drawbacks. Numerous researchers have shown that these algorithms fail to match human performance when slight disturbances are added to the data, see for example, Dodge et al. that illustrates that humans perform better than a DNN when there are image quality distortions Dodge and Karam (2019). In another paper, Pei et al. (2017) systematically explore limitations of DL systems by changing lighting and adding various occlusions to images. They find thousands of errors across their test program. These challenges of working with AI based in statistical learning motivate the need for AI Assurance.

In this manuscript, we present a blueprint for the assurance of AI systems. The proposed infrastructure aims to increase adoption of AI systems through increased trustworthiness, explainability Gunning et al. (2019), safety, and other assurance goals Cantero Gamito and Ebers (2021) (see Table 1 for definitions). Additionally, we advise the AI engineer to consider factors and challenges that influence assurance, such as model quality and data management Hossin and Sulaiman (2015). Finally, we propose a six-step generic process for applying assurance to any AI deployment.

Table 1. AI Summary Terms.

Term	Definition	Reference
AI Assurance	A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users.	Batarseh et al. (2021)
AI Domain	The organizational mission, domain (such as healthcare, economics, and energy), and associated systems/requirements pertaining to the AI enabled system.	Gunning et al. (2019)
Bias	The case when an algorithm produces results that are systemically prejudiced due to erroneous assumptions or inputs.	Nelson (2019)
Causality	The underlying web of causes of a behavior or event and furnishes critical insights that predictive models fail to provide	Pearl (2009); Cantero Gamito and Ebers (2021)
Data Democracy	Making digital information (i.e., data) accessible to the average non-technical user of information systems, without having to require the involvement of IT	Batarseh and Yang (2020)
Domain Dependence	Aligning an AI algorithm's utility with technical capabilities, industry-specific systems, and related requirements	Gunning et al. (2019)
Ethicality	An AI algorithm's ability to incorporate moral judgements based on right vs. wrong, morality, and social responsibility	Coeckelbergh (2020)
Explainability (XAI)	An AI system that is developed with high transparency and in a manner that promotes laymen level understanding of its algorithm and rationale.	Gunning et al. (2019)
Fairness	An AI algorithm's ability to ensure that an outpu reflects the whole population and its demographics.	Pearl (2009); Cantero Gamito and Ebers (2021)
Model Quality	(Re)-Training and (Re)-Testing of a model to optimize model metrics like accuracy and precision to iteratively and continuously improve a model's predictive power.	Hossin and Sulaiman (2015); Santhanam (2020); Rushby (1988)
Model Dependency	Set of libraries, code, and capabilities necessary for an algorithm to run.	Pearl (2009); Batarseh and Yang (2020); Cantero Gamito and Ebers (2021)
Operating Envelope	Envelopes are directly connected to the environment in which models run. They concern the external factors that impact data acquisition that affect model operation, training, testing, and execution through direct or indirect interactions.	Batarseh et al. (2021); Cantero Gamito and Ebers (2021)
Reliability	The removal of bugs, faults, and intrinsic errors in a model to enable its predictions to be consistent over time.	Cantero Gamito and Ebers (2021); Batarseh et al. (2021)
Robustness	The efficacy of a model to scale to other similar but different data sets and produce consistent results.	Cantero Gamito and Ebers (2021); Batarseh et al. (2021)
Transparency	Stating outputs and decisions of AI in a manner that can be explained in understandable (and preferably domain-specific) terms and formats to facilitate improved understanding of safety and compliance goals	Samek et al. (2019); Coeckelbergh (2020)
Trustworthiness	Confidence that a decision provided by an AI algorithm is reliable and would pass the Turing test in that it could be the same outcome created by a human user; which leads to trust.	Batarseh et al. (2021); Banks and Ashmore (2019)

What Is AI Assurance?

While there is extensive literature defining AI, the field of AI assurance is relatively new. While, researchers, technologists, scientists, policymakers, domain experts and business leaders have used numerous terms to describe the concepts related to assurance, there is no consensus on what this term precisely refers to. Batarseh, Freeman, and Huang define AI Assurance as, "A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users" Batarseh et al. (2021). This paper aims to serve as a foundational introduction to AI Assurance, its present status, rising need for it, existing methods, and future challenges. We provide an initial framework for the processes, metrics, and tools that should be associated with an AI assurance program.

In defining AI assurance, we leverage the above definition as well as definitions from both software assurance and quality assurance. Software assurance is defined as "the level of confidence that software functions only as intended and is free of vulnerabilities, either intentionally or unintentionally designed or inserted as part of the software, throughout the life cycle" NDAA (2013). According to the American Society for Quality (ASQ), quality assurance is the "part of quality management focused on providing confidence that quality requirements will be fulfilled." An alternative definition used by ASQ that will be helpful in defining AI Assurance is quality assurance is "all the planned and systematic activities that can be demonstrated to provide confidence that a product or service will fulfill requirements for quality." ASQ highlights that the confidence provided by quality assurance can be leveraged internally to management and externally to customers, government agencies, regulators, certifiers, and third parties.

In accord with these definitions, we define AI Assurance as *the probability that a system leveraging AI algorithms functions only as intended and is free of vulnerabilities throughout the life cycle of the system* Batarseh et al. (2021). High levels of assurance stem from all the planned and systematic activities applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy, and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users. A point of emphasis is that we focus on *systems* leveraging AI, the implications are that algorithm performance should be characterized relative to the deployed systems, and that assurance should reflect the deployed algorithm environment. Additionally, we emphasize that AI assurance stems from the planned systematic activities that span the system life-cycle. Clearly, we are leveraging the fields of software and quality assurance in defining how to achieve AI assurance. However, assuring AI has some striking differences that suggest it be treated as its own discipline. We emphasize that AI assurance is a multi-dimensional assessment—the probability that a system functions as intended includes assessments across constructs to include trustworthiness, explainability Samek et al. (2019); fairness Coeckelbergh (2020); Pearl (2009), assessment of unintended biases Nelson (2019), and the assessment of ethical implications Coeckelbergh (2020).

This paper discusses the "planned and systematic activities" needed for assuring AI. Included in this is the validation and integration of AI into production systems. This requires a unique understanding of the elements required to deliver repeatable and consistent predictions across numerous performance constructs via any set of deployed models. We outline a framework for AI Assurance that provides the information to make risk-based decisions on the adoption of AI into critical systems. Our goal is to provide information such that organizations and policy makers can develop processes enabling them to deploy AI across multiple domains and be confident it is free from vulnerabilities and will perform in a stable consistent manner.

The organization of this paper is as follows. The section that follows describes the large scale policy initiatives serving as motivation for AI Assurance in enabling wide scale AI adoption. We next introduce terminology and commonplace definitions that we will leverage (see Table 1). The next section will discuss the factors that impact changes in AI predictive power and performance. The next section describes a proposed framework for the "planned and systematic activities" that will contribute to an AI Assurance program. We conclude with challenges and recommendations for future work.

2. Motivation

AI strategy has become an international priority, the United States, Japan, the United Kingdom, Russia, Germany and China are just a few of the dozens of countries that have issued national strategies around AI, see for example Allen (2019); NAT (2019). In the USA, as agencies have thought through the use of AI in their agencies missions they have developed implementation strategies, guiding principles, and ethics statements. In 2019, the U.S. AI Initiative summarized American AI values in four different aspects: Understandable and Trustworthy AI, Robust and Safe AI, Workforce Impact, and International Leadership. The Department of Defense AI strategy emphasized leveraging AI for key missions, the democratization of data and capabilities through common analysis platforms, workforce development and recruiting, wide-scale collaboration, and highlighted that the Department of Defense (DoD) must lead in ethics and AI safety DoD (2018). The National Institute of Standards highlights that AI provides the opportunity to revolutionize metrology at the National Institutes for Science and Techhology (NIST) and emphasizes the importance of trustworthy, objective, explainable, and reliable AI. The Department of Energy notes the opportunity for AI to impact scientific data analysis, enhance modeling and simulation, and improve management and control of complex systems. They also highlight that AI needs to be interpretable, robust, and domain aware. The Intelligence Community developed a principles document emphasizing integrity, objectivity and equality, transparency and accountability, security and resilience, and that AI should be both informed by science and technology, and have a human-centered development and use EXE (2019); ODN (2019).

In reviewing these national strategy documents clear patterns emerge:

- Mission Alignment—each of these agencies emphasizes how AI will contribute to the agency's core mission.
- Informed by scientific understanding—using terms like domain aware, informed by science and technology, these documents highlight that AI algorithms should incorporate existing knowledge.
- Explainability—using terms like interpretability, transparency, and accountability, these documents highlight how important it is to explain AI decisions and predictions Samek et al. (2019).
- Fairness—terms such as objective, accurate, and equitable highlight how we need to understand any biases resident in AI algorithms, even if they are capturing and reflecting existing biases in society Pearl (2009); Coeckelbergh (2020).
- Reliability, Robustness—highlight the need to understand the consistency of algorithm prediction across varying domains and understand when they may fail Cantero Gamito and Ebers (2021).
- Secure—Algorithms and data should be protected and controlled at the appropriate level, algorithms need to be robust to adversarial action NIST (2020).
- Ethical—as AI scales solutions, it also scales mistakes, discrimination, and potential non-ethical outcomes. Algorithms ought to be assured to ensure ethical standards (within the context it is applied) are met Coeckelbergh (2020).
- Trustworthiness—concepts such as integrity, human-centered development and use, respect the law, convey that humans must be able to trust the AI algorithms for wide-scale (agency level) adoption Banks and Ashmore (2019).

Government agencies are clearly setting up the context for what information they will need to enable the widescale adoption of AI at an agency level. Similarly, Santhanam (2020) highlights the relatively low adoption of AI in business critical application, citing unique concerns around AI quality metrics Hossin and Sulaiman (2015) (e.g., explainability Samek et al. (2019), fairness Coeckelbergh (2020), reliability Cantero Gamito and Ebers (2021), etc.) as one of the challenges that need to be overcome. Santhanam (2020) notes, "As the application of AI moves to business/mission critical tasks with more severe consequences, the need for a rigorous quality management framework becomes critical." We concur with that assessment and expand to note that as government agency adoption will require such a framework as their missions can have wide scale societal implications. The ability to assess an algorithm across all of these dimensions provides motivation for an assurance framework for AI. That could enable AI assurance to be standardized on a global scale.

3. AI Assurance Definitions and Main Terms

Assurance is at the intersection of four points that are introduced in this section.

1. The difference between validation and verification: In the software testing world, testing could be categorized into two groups: validation and verification (often referred to as V&V). In simple terms, validation means providing the desired system to the user, it is building *the right system*, while verification is building the *system right* (i.e., without any errors, bugs, or technical issues). A conventional software system however, doesn't *learn*, it is based on predefined and preexisting sets of commands and instructions. AI assurance requires V&V, but it certainly expands beyond those limits. One of AI assurance's aspects that is fairly novel is Explainable AI (AI) Samek et al. (2019); Batarseh et al. (2021).
2. Test and Evaluation: Chapter 8 of the Defense Acquisition Guidebook defines the purpose of a Test and Evaluation (T&E) program is to provide "engineers and decision-makers with knowledge to assist in managing risks, measure technical progress, and characterize operational effectiveness, operational suitability, and survivability (including cybersecurity), or lethality of the system in the intended operational environment." To achieve this goal, a T&E programs should use a structured sequence of tests across the development life-cycle of a system coupled with statistical analyses of the data. Traditionally, T&E programs have ended at system fielding. However, the increase in software defined systems has pushed the need for ongoing T&E on fielded systems, the incorporation of AI software will further exacerbate this need. In this paper, we will look at T&E as the data collection process for achieving system validation for AI Assurance.
3. Explainable AI (XAI): References the concept of explainability Samek et al. (2019) at the intersection of multiple areas of active research. We discuss the following aspects of XAI:
    (a) Transparency: users of AI systems "have a right" to have outputs and decisions affecting them explained in understandable (and preferably domain-specific) terms and formats. That allows them to inspect its safety and clear goals.
    (b) Causality: besides correct inferences and predictions, an explanation for the underlying phenomena (i.e., data patterns or neural network layers) would be very helpful in increasing the trust in the system, and therefore, in its outputs (which affects domain-relevant assurance Pearl (2009).
    (c) Bias: bias has two forms in the context of AI model building. It can be statistical, and could be detected through overfitting and underfitting measure (which is opposite to variance). Bias can be also due to issues such as skew or data incompleteness in "the environment"; this aspect could be investigated and mitigated through data collection best practices, and the analysis of contextual awareness Nelson (2019).
    (d) Fairness: If decisions are made based on an AI system, they need to be verified that they are "fair". The infamous story of Google's racist bot is an example of what should be avoided Pearl (2009); Cantero Gamito and Ebers (2021).
4. Data democracy and context: The results of data science endeavors are majorly driven by data quality Santhanam (2020); Hossin and Sulaiman (2015). Throughout these deployments, serious show-stopper problems are still generally unresolved by the field, those include: data collection ambiguities, data imbalance and availability, the

lack of domain information, and data incompleteness. All those aspects are at the heart of AI assurance. Moreover, context plays a pivotal role in data collection and decision making as it can change the meaning of concepts present in a dataset. The availability of data at an organization to the *data collector* is an example of *democratization*. However, the *scope* of data required to create the desired outputs is refereed to as the *context* of the data collection (i.e., how much data is enough?).

5. AI subareas: In this document, we aim to establish subareas of AI clearly. As stated, different areas require different assurance aspects, and so we aim to capture those categories: a. Machine Learning (including supervised and unsupervised learning), b. Computer Vision, c. Reinforcement Learning, d. Deep Learning (including neural networks), e. Agent-Based Systems f. Natural Language Processing (including text mining), g. Knowledge-Based Systems (including expert systems).

4. Factors Influencing AI Assurance

AI is often assessed by its ability to consistently deliver accurate predictions of behavior in a system. A critical, often overlooked, aspect of developing AI algorithms is that performance is a function of the task the algorithm is assigned, the domain over which the algorithm is intended to operate, and changes to these elements in time. These parameters and their constituent parts form the basis over which assuring AI becomes a challenge. Algorithms need to be characterized by understanding the factors that contribute to stable performance across an operational environment (e.g., no dramatic perturbation by small changes and/or no effects measurable over time) Rushby (1988); Jha et al. (2020).

In order to accurately and consistently predict behaviors in systems, AI software requires data for training, validating, and testing predictions. The iterative process of improving accuracy and precision in developed models involves trade-offs in performance, data quality, and other environmental factors Santhanam (2020). AI's predictive power can be impacted through changes in the training and test data, the model, and the environment. In this paper we will discuss sources of change captured within an operational envelope Batarseh et al. (2021) of an AI's execution that is often attributed to inconsistencies in AI software. Model and data changes have been discussed in the literature around concept drift Žliobaitė (2009); Gama et al. (2004); Gama et al. (2014); Tsymbal et al. (2006); Tsymbal (2004) and are examples of how these inconsistencies could be measured.

4.1. Data Democracy and Quality

Data (besides the algorithm) are the main fuel for any AI model. As global connections continue to improve, more people are finding themselves living in a "data republic." In this republic, data are generated, collected, stored, and analyzed. Whoever owns the data, will have a ruling centralized hand over others, which would lead to an absolute need for the democratization of data (i.e., decentralization) Batarseh and Yang (2020). For instance, data openness at governments is critical to increasing transparency and accountability, but most importantly democratized data can allow for an open discussion on policies and all aspects of decision making in a democratized manner. In a democratized data republic, citizens of an organization or a country have both rights and responsibilities to shape data usage in ways they most value.

4.2. Notion of an Operating Envelope

A core concept for assurance is understanding the domain over which an AI algorithm can be expected to perform based on the data it has been trained on; this is called the operating envelope of a model Batarseh et al. (2021). A model's operating envelope describes the environment and its factors that impact the observations and measurements (i.e., the data) gathered prior to building a model. The envelope directly impacts model performance and operation. An example, can be taken from examining the close correlation between the time of day, shadows, and observation angles in gathering overhead satellite imagery in the development of computer vision models.

Envelopes are directly connected to the environment in which models run. They concern the external factors that impact data acquisition that affect model operation, training, testing, and execution through direct or indirect interactions. Users and their operation of models and/or interpretation can be included in these considerations. Pertaining to AI and ML, the character of how measurements and observations are recorded in the form of data, that is environmental considerations, directly impact the quality of models and their subsequent assurance. As observations are recorded and models are trained over this collected data, change may be introduced into the model as it is re-trained and re-tested that may not be accounted for. This is the notion of concept drift Gama et al. (2014); Tsymbal et al. (2006); Gama et al. (2004), Žliobaitė (2009). As AI and ML are specifically purpose built, context specific, environmentally dependent, and envelope sensitive, quality and assurance may be adversely impacted Kläs and Vollmer (2018) as a function of the derived concept drift from any of these factors. Based on this, one can expand the definition of AI assurance as the confidence range of AI and ML to to deliver consistent and reliable predictions in connection and correlation with an operating envelope Cantero Gamito and Ebers (2021); Batarseh et al. (2021).

4.3. Hardware Design Model Integration

Hardware can greatly impact the the frequency of (re)-training, (re)-testing, and (cross)-validation. Model operations, i.e., devsecops practices applied to ML, have optimized the use of available compute with automation of relevant pipeline tasks pertaining to AI/ML development lifecycles. Related to this are scenarios where the existence and subsequent application of hardware to increasing compute footprints can provide optimizations to address factors related to drift. In this case the usage of hardware can greatly influence the frequency of (re)-training and (re)-testing thereby improving overall model performance.

4.4. Statistical Considerations

Model performance is effected by two main components: the data, and the algorithm selected. Data measures such as bias, sampling, skew and others are to be accounted for before the algorithm is trained. Incomplete data, biased data, or unclean data would lead to incomplete results, biased results, and unclean results (i.e., garbage in, garbage out—GIGO). Accordingly, data science best practices need to be implemented. For instance, the equilibrium between bias and variance is critical in avoiding any over or underfitting Rushby (1988); Santhanam (2020), a key concept in intelligent algorithms. For example, if we consider the relative standard error (RSE) of a statistic (defined as the standard error of the estimated statistic divided by the estimated statistic), the strictness of the RSE measure varies depending on the level of the estimated proportion (i.e., in sampling), RSE can be too conservative for very small proportions, but too liberal for very large proportions.

Moreover, the selection of a model is another hurdle that needs to be managed through a deliberate and educated process. For instance, if a classification problem is presented, there are multiple algorithms that could “do the job”; however, in most cases, only few algorithms provide useful, actionable, and explainable results. In ML for example, boosting algorithm proved successful because they aggregate the best results from multiple executions of the model with different parameter setups. That notion is referred to as Ensemble ML (EML), and it reduces the need of “manually” testing multiple executions Jha et al. (2020). During model execution, if three executions of a classification algorithm are performed, the outcome of all three models yields better (more accurate classifications) than any of them independently.

Those two examples illustrate the need for a careful and exact process while handling data, as well as defining and executing the algorithm. However, such practices are also effected by the domain.

5. Proposed Framework for Assuring AI

In developing a process for assuring AI, we will leverage the planned systematic activities that support quality assurance and software assurance. Six Sigma is arguably the most well known and widely used quality management methodology. The DMAIC (Define, Measure, Analyze, Improve, and Control) is an integral part of the Six Sigma methodology Shankar (2009). We will focus on DMAIC as it focused on the acquisition of data through purposeful planning, measurement, testing, and analysis. A quick review of Google Scholar provides insight into how DMAIC has been applied from everything from yogurt making to power-plant engineering.

Inspiration for assuring AI can also be drawn from software assurance. In general, consistent and stable performance of application software and its security can be framed based on gathered metrics on performance, vulnerabilities, and weaknesses. Risk assessments of software deployed on systems utilize the risk management framework (RMF) NIST (2020) assuring software on government systems. Risk based findings based on these metrics (e.g., existence vulnerabilities) are typically required to be mitigated. In case these findings cannot be mitigated, a plan of action and milestones (POAM) is necessary to provide the approach to address the risk. Statistical learning based AI can be considered as ′stochastic software′ whose assurance involves characterizing its risk of performance and vulnerability on a system over time.

Similar to DMAIC, the RMF follows and iterative process: Categorize Information System, Select Security Controls, Implement Security Controls, Assess Security Controls, Authorize Information System, Monitor Security Controls.

These frameworks suggest an underlying process to assuring quality in complex systems. A general process they capture is (1) define the problem/goal; (2) provide structure by defining measurement; (3) understand context; (4) developing a data collection and analysis approach (e.g., T&E strategy); (5) execute the strategy; (6) making improvements based on information gained; (7) and finally controlling/monitoring the systems throughout use. The International Statistical Engineering Association have codified this process in Figure 1.

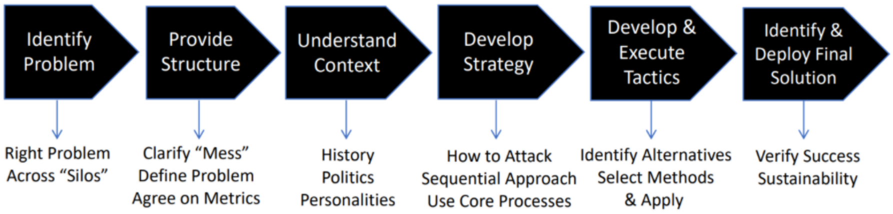

Figure 1. Typical Statistical Engineering Process, adapted from Hoerl and Snee (2020).

We will discuss the specific realizations of this process that are critical for AI Assurance below.

5.1. Define

The define stage sets up the foundation for the AI assurance process. This phase should focus on ensuring that the AI assurance program is adequate. A key aspect of the define phases is ensuring that all stakeholders are involved in the defining process. Stakeholders may include the mission/task champion (leadership), program management, system engineer, AI developer, requirements representative, test and evaluation personnel, end users, etc. depending on the application. Based on an extensive literature review Batarseh, Freeman, and Huang Batarseh et al. (2021) identify three elements of AI assurance that are important in the "Define" phase: AI Subarea, AI Domain, and AI Goals, we expand on that list here. Critical elements to define in this phase include:

1. AI Domain—what is the broad organization mission that is acquiring the AI enabled system (e.g., government, energy sector, defense, healthcare, etc.)?; What context does that bring to the challenge of assuring AI? Gunning et al. (2019)
2. Mission context—what is the problem that the organization is trying to solve with an AI system(s)? How does the AI support the organizational mission? Gunning et al. (2019)
3. AI Area—what is the type of AI needed (e.g., deep learning, reinforcement learning, etc.); How does the AI contribute in ways that previous methods failed?
4. Scientific/Engineering Alignment—What are the scientific and/or engineering needs that the AI can solve?; How does it integrate with know constraints?; How does it incorporate prior knowledge (solely through data or other mechanisms)?
5. AI Goal—What are the primary directives for the AI, how does this stem from the AI domain? Must it be ethical, explainable, fair, etc.? Pearl (2009); Cantero Gamito and Ebers (2021)

In this phase illustrative user stories or use cases can be helpful in connecting the domain, mission context, AI solutions, how prior information and AI goals. They also provide context for the measurement of success and failure.

5.2. Measure

In this paper we have touched on numerous dimensions of AI measurement, not all intelligent systems will require all dimensions. The measures should stem from the problem definition. Dimensions of measurements that should be considered include:

1. Algorithm Performance Cantero Gamito and Ebers (2021)
2. Bias/Fairness Nelson (2019)
3. Security Banks and Ashmore (2019)
4. Safety Batarseh et al. (2021); Banks and Ashmore (2019)
5. Trustworthiness Cantero Gamito and Ebers (2021)
6. Explainability Samek et al. (2019)
7. Ethicality Coeckelbergh (2020)

Although those measures are highly subjective, methods to quantify them are needed to further advancing the field. The loading of ethical standards or trustworthiness Banks and Ashmore (2019) for instance into an AI system is -in most cases- domain specific; for instance, an ethical metric in a system built for a healthcare application will be -most likely- different than one in a warfare application, therefore, the process of "values loading" into the AI system is needed to capture those measures and create a benchmark to compare against for assurance quantification. In this paper, we don't propose a quantification method, rather, we point to all the aspects required for consideration in that regard. Additionally, traditional measurements of systems should be considered to include system performance, utility, efficacy, usability, resilience, reliability, maintainability, among others Hossin and Sulaiman (2015). These measures when coupled with AI specific measures provide a holistic assessment of not only the dimensions of assurance, but also their correlations with system outcomes.

What measurements are needed and how to best measure is a complex task in itself Bloomfield and Rushby (2020). In addition to selecting the right measures, one will need to consider implications of data quality and accessibility and to what degree algorithm specific measurements are needed Cantero Gamito and Ebers (2021). "Black box" algorithms may require additional measures or more frequent measurement to account for the lack of algorithm transparency.

5.3. Characterize & Add Context

The measurement of intelligent systems leveraging AI is dependent on the multidimensional space that includes the operational environment, system, hardware infrastructure, and other factors effecting the AI. Siebert et al. (2020) in their paper, "Towards Guidelines for Assessing Qualities of Machine Learning System" provide several context lenses that should be considered, the model view, data view, infrastructure view, and environmental view. These views illustrated in Figure 2. Siebert et al. focus on specific quality measurements for each view, which should be considered as part of the definition of measurement.

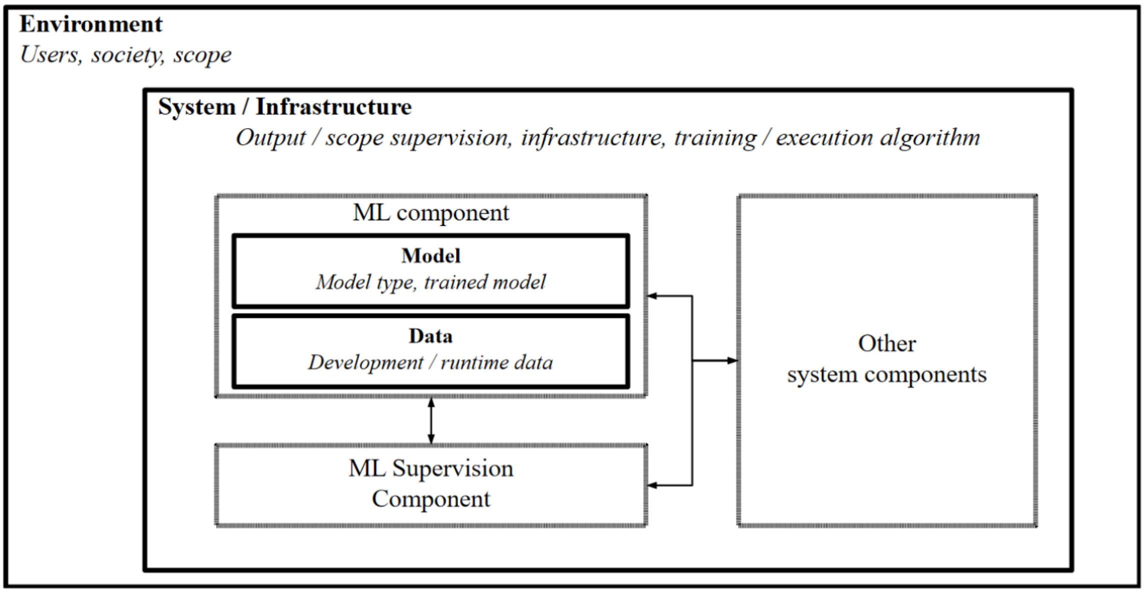
Environment
Users, society, scope
System / Infrastructure
Output / scope supervision, infrastructure, training / execution algorithm
ML component
Model
Model type, trained model
Data
Development / runtime data
ML Supervision Component
Other system components

Figure 2. Overall Environment and AI System Siebert et al. (2020).

The different "views" provide a useful context for thinking about what aspects of the system need to be characterized for assurance. The concept of an operating envelope applies at each view level. For example at the operational environment view one might be concerned about factors that include environmental conditions, operational users, mission tasks, and other factors external to the system impact performance outcomes. At the lower level the operating envelope many consist of data views like data quality, information content, domain, and range. Each of these views needs to be considered for their impact on achieving the goals of the intelligent system and appropriately planned for in the execution of the assurance program.

5.4. Plan Strategy

We emphasize the need for leveraging test and evaluation across the system development, deployment, and operations life-cycle as the basis for a strategy for assuring AI. Santhanam (2020) provides Figure 3 to highlight the artifacts in the development of a machine learning algorithm. He emphasizes the need for iterative processes and several quality improvement tasks that include multiple types of testing, but also debugging, manual inspection, and static analysis. While formal verification strategies can be a component of an assurance strategy, the complexity of statistical learning algorithms—especially when considered as employed as part of a system that interacts with the operational environment—demands a data driven approach to assurance. The planning strategy should consider the use of formal verification approaches as well as test venues that enable the exploration of AI model outputs directly, the incorporation of digital simulations, software-in-the-loop simulations, hardware-in-the loops simulations, full system testing in controlled environments, and full system testing in operational environments with operational users.

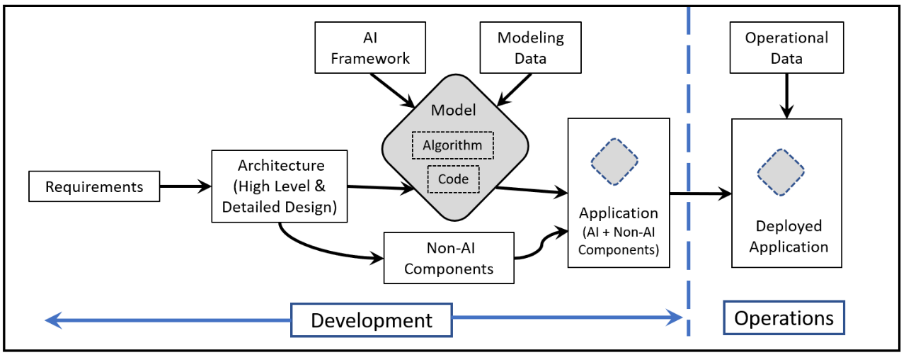

Figure 3. Artifacts in a Quality Management Framework Santhanam (2020).

The planning strategy should describe the iterative, sequential process for gathering knowledge across the key measurements for AI assurance. Testing the system should occur across the developmental and engineering life-cycle. Statistical methods such as design of experiments, sequential-adaptive test design techniques (e.g., response surface methods, optimal learning), and stratified random sampling for algorithms training, testing, should be employed to provide a systematic approach to scaling information collection with system development and ensuring adequate information exists across the various views on operating envelopes exists to support the assurance assessment.

5.5. Execute & Analyze

The execution phase of testing should also include ongoing analyses of the data, assessments of feasibility, and lessons learned. Analyses should consider knowledge gaps based on planning strategies and dedicated time for reworking the test program based on information obtained during each phase of testing.

5.6. Monitor & Improve

In their paper, "Test and Evaluation Framework for Multi-Agent Systems of Autonomous Intelligent Agents" Lanus et al. describe a continuous process of T&E for autonomous intelligent agents that shows that this data collection process is followed throughout development, manufacturing, deployment, and operations (as shown in Figure 4). In operations, they emphasize the need for both ongoing monitoring of the systems capabilities, independent assessment when pre-defined triggering events such as a mission objective change or an adversarial attack. Additionally, software defined functionality in general provides the opportunity to improve capabilities without a full redesign of the system. AI algorithms can be improved by introducing new more relevant data for the current tasking. Methods such as transfer learning and targeted fine-tuning provide the opportunity for continual improvement to fielded algorithms. However, robustness is always a consideration that must be considered when making algorithm improvements.

The monitor phase is essential to consider from the early phases of system design. Monitoring is an easier task when the system is designed to output relevant measures and metrics automatically. Similar to a gas monitor on a physical systems, AI algorithms need "health" monitoring built in from the system origin. For data driven monitoring statistical quality control and control charts in general provide a mechanism for understanding if variations in performance are due simply to the stochastic nature of AI algorithms deployed in operational environments or a unexpected departure from previously observed performance.

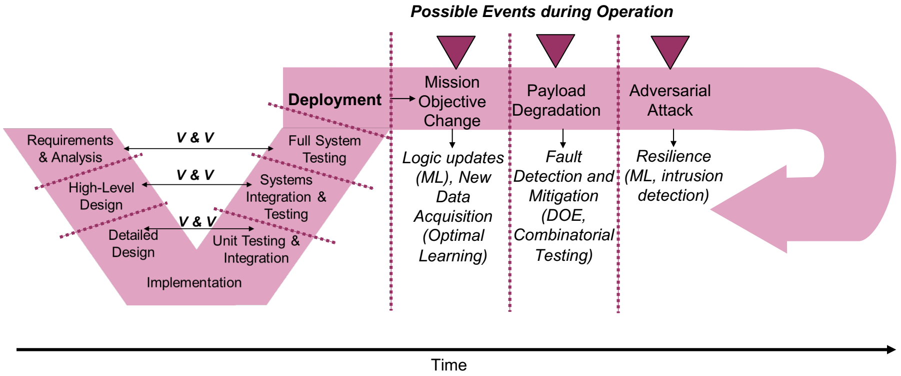
Possible Events during Operation
Deployment
Mission Objective Change
Payload Degradation
Adversarial Attack
Logic updates (ML), New Data Acquisition (Optimal Learning)
Fault Detection and Mitigation (DOE, Combinatorial Testing)
Resilience (ML, intrusion detection)
Requirements & Analysis
High-Level Design
Detailed Design
Implementation
Unit Testing & Integration
Systems Integration & Testing
Full System Testing
V & V
V & V
V & V
Time

Figure 4. The VTP Model from Lanus et al. (2021) extends the Systems Engineering “Vee” model Deployment and Operations.

6. Challenges in Assuring AI

The research and development of AI-driven methods is consistently yielding better results in solving complex problems across multiple domains. However, these advancements only stress the need for more assurance. The research community is still scratching the surface when it comes to claiming victory on valid and verified intelligent systems, mostly due to the deployment challenges mentioned in this section. We present six major challenges:

1. Model Quality. in Siebert et al. (2020) discuss the importance of quality metrics in reference to AI and ML. The primary driver of their work was to motivate the need to develop good quantitative measures of quality to connect context to function in an operational setting. A critical arena for this correlation is in the consideration of what is called the ground truth of the model. Existence of ground truth can occur in three modes: full, partial, or not at all. In each of the three scenarios, its existence can directly determine the ability to effectively determine quality measures. In the case of full ground truth being known, the ability to characterize quality can be done in direct measure with the known ground truth. In the case of partial ground truth awareness, considerations of data quality must be examined meticulously. Finally, if no ground truth is known the training and test splits are based on assumptions that require delicate evaluation as it pertains to the quality of model predictions. In all three cases, it becomes clear that metrics associated with the characterization of model performance should be used to proffer connections to the quality of model outputs (i.e., predictions) Hossin and Sulaiman (2015).
2. Usage of Model Metrics. Model metrics differ based on the types of models being developed. Classification models utilize accuracy, precision, f-score, and others while clustering methods Emmons et al. (2016) utilize others such as Silhouette and Elbow Diagrams. In evaluating the quality of models, the treatment of these metrics as it pertains to determination of false positives (FP) and false negatives (FN) may be weighted differently as its applied to the specific model. Compatibility considerations for FP and FN have also been shown to play a part in forming deeper understandings of quality and assurance practices Banks and Ashmore (2019); Samek et al. (2019).
3. Model Dependency. Furthermore, while supervised ML models have pre-labeled outcomes (i.e., predictions) that could be verified against actual numbers, unsupervised models don’t have the same labels as the outcomes are dependent on the patterns found in the data, therefore, AI assurance is *model-dependent* Batarseh et al. (2021).
4. Domain Dependence. Additionally, different domains have different “expectations”, for instance, a 0.1 variance in revenue predictions for decision making at a company has much more benign consequences than a 0.1 variance in an open heart surgery robot or a mission-critical bomber aircraft. Therefore, AI assurance measures are *domain-dependent* Gunning et al. (2019).

5. Geography. Besides the "domain" challenge, there is a global (geographical) issue based on environmental, geographic, and geospatial (and even cultural) factors that contribute to (re)-training, (re)-testing, and (cross)-validation of models. For example, as weather patterns may alter due to seasons or other factors, ML performance may vary likewise. Understanding the impacts of these changes are critical to model stability.
6. Operating Envelopes. Minimization of factors contributing to operating envelope change is important. The observations and measurements (i.e., the data) gathered prior to building a model should be stable over time. Challenges related to maintaining this stability can drastically impact model performance and operation Batarseh et al. (2021).

7. Conclusions, Summaries, and Future Work

If assurance and testing in AI is inspected, it is evident that prior to the big data revolution, handcrafted knowledge was king. AI was in the form of expert systems, heuristic reasoning, inference, and other forms of knowledge-based deployments. In the 1980s and 1990s, AI experienced what is referred to as the AI winter. During that phase, research slowed down, funding froze, and scientists turned their eyes to other areas (such as web development, software engineering, and quality). However, in the early 2000s, statistical learning evolved further and became possible due to the availability of data. All forms of learning, such as ML, DL and RL redefined (and rekindled) the field of AI. The new wave of statistical learning based-AI has created the need for new methods of AI assurance. A summary of AI-relevant terms is presented in Table 1.

In this paper we have provided an overview of the different dimensions of AI assurance for statistical learning based AI (see Table 1 with corresponding definitions). We argue that due to the stochastic nature of statistical learning based AI new processes combining the best of past methods of AI Assurance, Software Assurance, and Quality Assurance are needed. In the future notions such as contextual adaptations, causality, and XAI will become more evident and dominant. Nonetheless, one aspect that all phases of AI require is assurance. One aspect is key: For AI to reach its scientific and practical goals, and for humans to reap its benefits, AI researchers, practitioners, and investors are on a mission to display the virtuous goodness of a fair, assured, and explainable AI. If that notion is not clearly deployed and demonstrated, AI will surely experience another winter.

Author Contributions: L.F. designed the overall structure of the paper and developed the proposed AI assurance framework; A.R. wrote multiple parts of the paper such as "hardware design integration", and helped with the overall paper vision; F.A.B. wrote different parts of the paper, defined the main terms, and developed the factors influencing AI assurance. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

Allen, Gregory C. 2019. Understanding China's Ai Strategy. Available online: https://www.cnas.org/publications/reports/ (accessed on 30 April 2021).

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, and et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. Paper presented at the International Conference on Machine Learning, New York, NY, USA, June 19–24; pp. 173–82.

Banks, Alec, and Rob Ashmore. 2019. Requirements assurance in machine learning. Paper presented at the AAAI Workshop on Artificial Intelligence Safety 2019, Honolulu, HI, USA, January 27; pp. 14–21.

Batarseh, Feras A., and Ruixin Yang. 2020. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Systems*. Cambridge: Academic Press.

Batarseh, Feras A., Laura Freeman, and Chih-Hao Huang. 2021. A survey on artificial intelligence assurance. *Journal of Big Data* 8: 1–30. [CrossRef]

Bloomfield, Robin, and John Rushby. 2020. Assurance 2.0: A manifesto. *arXiv*, arXiv:2004.10474.

Cantero Gamito, Marta, and Martin Ebers. 2021. *Algorithmic Governance and Governance of Algorithms: An Introduction*. Cham: Springer International Publishing, pp. 1–22.

Coeckelbergh, Mark. 2020. *AI Ethics*. MIT Press Essential Knowledge Series. Cambridge: MIT Press.

Department of Defense. 2018. *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*. Report. Washington, DC: Department of Defense.

Dodge, Samuel, and Lina Karam. 2019. Human and dnn classification performance on images with quality distortions: A comparative study. *ACM Transactions on Applied Perception (TAP)* 16: 1–17. [CrossRef]

Emmons, Scott, Stephen Kobourov, Mike Gallant, and Katy Börner. 2016. Analysis of network clustering algorithms and cluster quality metrics at scale. *PLoS ONE* 11: e0159161. [CrossRef] [PubMed]

Executive Order 13859: American Artificial Intelligence Initiative. 2019. Washington, DC: Executive Order, Executive Office of the President of the United States of America.

Gama, João, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. 2004. Learning with drift detection. In *Advances in Artificial Intelligence—SBIA 2004. Lecture Notes in Computer Science*. Berlin and Heidelberg: Springer, vol. 3171.

Gama, João, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* 46: 1–37. [CrossRef]

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—Explainable artificial intelligence. *Science Robotics* 4. [CrossRef] [PubMed]

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Paper presented at IEEE International Conference on Computer Vision, Santiago, Chile, December 7–13; pp. 1026–34.

Hoerl, Roger W., and Ronald D. Snee. 2020. *Statistical Thinking: Improving Business Performance*. Hoboken: John Wiley & Sons.

Hossin, Mohammad, and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5: 1–11.

Jha, Susmit, John Rushby, and Natarajan Shankar. 2020. Model-centered assurance for autonomous systems. In *Computer Safety, Reliability, and Security. SAFECOMP 2020. Lecture Notes in Computer Science*. Cham: Springer, vol. 12234.

Kläs, Michael, and Anna Maria Vollmer. 2018. Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In *Computer Safety, Reliability, and Security. SAFECOMP 2018. Lecture Notes in Computer Science*. Cham: Springer, vol. 11094.

Lanus, Erin, Ivan Hernandez, Adam Dachowicz, Laura Freeman, Melanie Grande, Andrew Lang, Jitesh H. Panchal, Anthony Patrick, and Scott Welch. 2021. Test and evaluation framework for multi-agent systems of autonomous intelligent agents. *arXiv*, arXiv:2101.10430.

Launchbury, John. 2017. A Darpa Perspective on Artifical Intelligence. Available online: https://https://www.darpa.mil/attachments/AIFull.pdf (accessed on 30 April 2021).

National Defense Authorization Act (NDAA) for Fy 2013. 2013. Bill Passed, United States. Congress. House. Washington, DC: Committee on Armed Services.

National Institute for Standards and Technology (NIST). 2020. *Security and Privacy Controls for Information Systems and Organizations*. NIST Special Publication 800-53, Rev. 5. Gaithersburg: National Institute for Standards and Technology (NIST).

Nelson, Gregory S. 2019. Bias in artificial intelligence. *North Carolina Medical Journal* 80: 220–22. [CrossRef] [PubMed]

Office for the Director of National Intelligence. 2019. *The Aim Initiative: A Strategy for Augmenting Intelligence Using Machines*. Technical Report. McLean: Office for the Director of National Intelligence.

Pearl, Judea. 2009. *Causality*, 2nd ed. Cambridge: Cambridge University Press.

Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. Paper presented at 26th Symposium on Operating Systems Principles, Shanghai, China, October 28; pp. 1–18.

Rushby, John. 1988. *Quality Measures and Assurance for Ai Software*. NASA Contractor Report 4187. Menlo Park: SRI International.

Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müllerüller. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science.* Berlin: Springer International Publishing.

Santhanam, P. 2020. Quality management of machine learning systems. In *Engineering Dependable and Secure Machine Learning Systems*. Cham: Springer International Publishing, pp. 1–13.

Shankar, Rama. 2009. *Process Improvement Using Six Sigma: A DMAIC Guide*. Milwaukee: Quality Press.

Siebert, Julien, Lisa Joeckel, Jens Heidrich, Koji Nakamichi, Kyoko Ohashi, Isao Namba, Rieko Yamamoto, and Mikio Aoyama. 2020. Towards guidelines for assessing qualities of machine learning systems. In *Quality of Information and Communications Technology. QUATIC 2020. Communications in Computer and Information Science*. Cham: Springer, vol. 1266.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529: 484–89. [CrossRef] [PubMed]

The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. 2019. Report by the Select Committee on Artificial Intelligence of The National Science and Technology Council, Executive Office of the President of the United States of America. Available online: https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf (accessed on 30 April 2021).

Tsymbal, Alexey. 2004. *The Problem of Concept Drift: Definitions and Related Work*. Technical Report. Dublin: Department of Computer Science, Trinity College.

Tsymbal, Alexey, Mykola Pechenizkiy, Padraig Cunningham, and Seppo Puuronen. 2006. Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections. Paper presented at 19th IEEE Symposium on Computer-Based Medical Systems (CBMS '06), Salt Lake City, UT, USA, June 22–23; pp. 679–84.

Žliobaitė, Indrė. 2009. Learning under concept drift: An overview. *arXiv*, arXiv:1010.4784.