Quality Characteristics of Artificially Intelligent Systems

Adam Leon Smith
Dragonfly
Email: adam@wearedragonfly.co

Raphaël Clifford
Email: raphael@clifford.net

***Abstract*—This paper explores the quality characteristics of systems using artificial intelligence components, referencing existing work in this area from research, the private sector and the international standards community. It proposes a new quality model that considers existing work, and where applicable, proposes measures that may be appropriate for each quality characteristic.**

***Index Terms*—artificial intelligence, machine learning, quality, systems engineering, quality management, testing**

I. Introduction

Artificial Intelligence (AI) can be loosely defined as applying acquired knowledge to make decisions, in contrast to using explicit logic. This presents both opportunities and problems [1] for the field of software and systems quality. One of the most challenging aspects is the the probabilistic nature of sub-symbolic systems, and difficulties in reproducing and explaining results. Another challenge is convincing everyone that AI systems can be trusted with important decisions, and some researchers propose that what AI systems actually need to achieve is reliability [2], that is related to quality.

Artificial intelligence can include symbolic rule-based expert knowledge systems and sub-symbolic (statistical) machine learning systems. Machine learning is the most common AI method, and it is difficult to specify quality, and analyse how to test. Research in Japan involving 278 machine learning engineers identified the biggest new challenges they face integrating machine learning is in decision making with customers and testing/quality assurance. Further, they identify the lack of a test oracle[1], and imperfection as the top causes of this [3].

Should a practitioner wish to define a strategy or approach to holistically evaluate the quality of an AI system, at present, it requires review of many scholarly articles in order to identify the relevant properties. It is therefore important that a standard quality model for AI systems is developed to support practitioners.

This paper reviews existing quality models in the context of AI, that is acquiring knowledge, applying the knowledge and producing decisions. Robustness and context completeness are introduced as characteristics that relate to the input domain; bias, functional correctness, and ex-post explainability (run transparency) as relating to the output decision domain; and adaptability, transparency, societal and ethical risk mitigation as non-functional characteristics. This paper gives examples of measures that can be used, but it not intended to be exhaustive on this matter.

[1]In software testing, a test oracle is a source to determine an expected result to compare with the actual result of the system under test [4]

II. Existing work

A. SQuaRE

ISO/IEC 25010 [5] is a popular standard in software and systems quality management defining system and software quality models, alongside ISO/IEC 25012 [6] that defines a data quality model. ISO/IEC 25010 has been identified by practitioners [7] as requiring modification for AI systems. ISO/IEC have also commenced a project to create a standardised model [8], that as an International Standard, consistent with ISO/IEC 25010, will drive efficiency in industry. This is expected to be published in 2023.

B. DIN Spec 92001-1

The DIN SPEC 92001-1 [9] is a freely available standard published in April 2019 by the German standardisation body (DIN). It aims to provide an outline of AI lifecycle process and quality requirements. It outlines three quality pillars: functionality and performance, robustness and comprehensibility. This paper refers to functionality and performance as functional correctness and completeness in order to stay consistent with existing ISO standards [5]. Similarly, this paper refers to comprehensibility as transparency. Nevertheless, the scope of the three quality pillars covered in the DIN SPEC are also covered in this proposed model.

C. ISO/IEC Standards on Artificial Intelligence

ISO/IEC are working on a large number of new standards relating to AI, including those that relate to quality, verification and validation. The first relevant standard to the quality topic is a technical report giving an overview of trustworthiness in AI [10]. This identifies vulnerabilities in AI systems including security and privacy threats, bias, unpredictability, opaqueness, and more. Some of these vulnerabilities map directly on the requirements for a quality model for AI.

III. Proposed quality model

A. Model Types

Quality cannot be quantified as a single metric. It requires the definition of characteristics and terminology

that can be used when specifying quality requirements, and evaluating them. ISO/IEC 25010 defines two models, a product quality model and quality-in-use model. The former is the characteristics of a static system, and the latter are characteristics of a system with a specified context of use. To draw an analogy, AI product quality (PQ) may be determined once at the point of release, AI quality in use (QiU) can only be determined in each actual context of use. In traditional software systems, few characteristics overlap. Table 1, below, lists the quality characteristics discussed in this paper.

Quality model in ISO/IEC 25010	*Characteristic*	*Discussed sub-characteristics*
Product quality	Functional suitability	Functional correctness
Product quality	Functional suitability	Bias
Product quality	Portability	Adaptability
Product quality	Security	Robustness to adversarial examples
Product quality	Usability	Run transparency
Product quality	Usability	Controllability
Product quality	Maintainability	Functional transparency
Quality in use	Context coverage	Context completeness
Quality in use	Freedom from risk	Societal and ethical risk mitigation

Table 1 - New quality sub-characteristics to be discussed

B. Functional Suitability

1) Functional Correctness: Whilst there are significant challenges relating to the verification of functional correctness and completeness, how to statistically analyse the results of classification and regression systems common in AI, is a mature topic. Measurement of type I (α) and type II (β) errors is one common approach [11] to presenting classifier results. For example, where H_0 represents all negative predictions and H_1 represents all type II errors (false negative predictions), the type II rate can be calculated as:

$$\beta = \frac{H_1}{H_0}$$

Regression problems also have various metrics, of which one of the most common is Mean Absolute Error (MAE), this gives no indication about the directionality of performance, simply the scale. Where Y_i represents predicted values, and X_i represents the ground truth, the error can be expressed as:

$$MAE = \frac{1}{n}\sum|X_i - Y_i|$$

2) Bias: Bias is a term that is frequently used differently by different stakeholders. It is common for data scientists to think of bias as a statistical property which can be positive or negative in any given context, and ethicists to think of bias an unfair outcome. In the context of an overall AI system bias is both, it is a property of data and an algorithm. Bias also manifests as cognitive biases that exist on the development team, and societal biases that exist in historical datasets.

Overall, bias in an AI system is a property of the system that results in different treatment for different people, objects or groups. In this context, it is an accuracy issue that exists in relation to the functional correctness and completeness of a system. Bias can be measured using MAE or α and β as described above, but in a way that filters out results for a particular cohort of transactions that belong to a specific group. In this way the results can be compared between the general population to identify bias. Another approach is to use statistical parity [12]. This uses S as a variable that identifies the cohort under analysis, and S = 1 indicates membership of the relevant cohort:

$$DI = \frac{P(Y = 1|S = 0)}{P(Y = 1|S = 1)}$$

Other metrics are required when assessing ranked outputs [13] or continuous variables, but the principle of comparing the group under analysis to the general population remains.

Bias belongs to both the product quality model (product bias), and the quality in use model (bias in use). This is because it is the property of a single system and the data inputs used in the production of that system, but it is also a property of the system in actual use, where the inputs may be very different.

C. Adaptability

Adaptability is defined as a product quality characteristic in ISO/IEC 25010:

> degree to which a product or system can effectively and efficiently be adapted for different or evolving hardware, software or other operational or usage environments

and is part of the portability characteristic - which refers to the environment. This paper proposes that the definition of adaptability is extended. It is much more the case with AI systems that the data observed by the system can now be part of the environment, in real-time with reinforcement learning, or as models are "retrained" with new datasets. This is starkly different to making a change to existing logic, as the model is completely re-baselined rather than incrementally changed, and the change may be interactive, dynamic, periodic or even in real-time.

Adaptability could be defined as the time taken for a system to start to react differently based on a change in observed data, or the ease with which it can be retrained.

D. Controllability

The degree to which a system can be controlled is not a new concept [34], and is typically a functional aspect of a system, however increasingly systems are able to operate without human intervention or control. Therefore, if human interaction becomes optional or impossible, it is important to consider how controllable an AI system is for its presumptive human operator. Controllability can be considered to be the ability to move a system from an arbitrary initial state, into

another desired state, by exercising a control function, within a certain number of steps and within the required time.

E. Robustness and adversarial examples

The environment in which AI must operate may be subject to change through natural perturbations and volatility, drift in the characteristics of input data over time or malicious intention of an adversary. The term AI Robustness attempts to capture the extent to which an AI system will safely maintain its intended behaviour in such scenarios. This is distinct from context completeness which does not focus on unanticipated changes in input distributions. Robustness is however to some extent captured under the catch-all term context coverage. Ensuring robustness poses some of the most difficult and important questions in AI and raises a number of issues which we will introduce below.

1) Distributional Change: Perhaps the most common challenging issue in AI is how to maintain the desired behaviour of a system when the input distribution changes over time. If the test data has the same statistical properties as the training set then we can expect a well specified AI system to work correctly. However, when encountering new and unexpected input the situation can be much more difficult. As an example, in "Concrete Problems in AI Safety" [16] an AI robot cleaner is posited which was trained to clean an empty office. On encountering a pet dog for the first time it might attempt to wash it giving unpredictable results. In a more serious settings such as when trading on the stock market or in military applications these consequences could be disastrous. In general, when the testing distribution differs from the training distribution AI systems might not only perform in unexpected ways but they may also report that they have been functioning without problems. This therefore makes the diagnosis of faults in the AI system problematic.

2) Adversarial Inputs: Attempts to fool AI systems date back at least 15 years to the early days of spam filters. Those wanting to send bulk unsolicited email started to find ways to avoid the linear classifiers used to filter them out. Since the resurgence of deep neural networks the importance of adversarial techniques has become of increasing interest and importance. It is now well known that computer vision systems can be fooled to make wildly inaccurate classifications if given a suitably perturbed image [17]. This failure of AI is in fact caused by an unanticipated distributional change in the input that was not captured in the training set. What sets this apart is that this difference has been carefully crafted to make the AI system give an incorrect response by a malicious adversary.

However these challenges are not just limited to AI based computer vision systems. Every year more and more classes of inputs, including malware detection [18] and natural language texts [19] are being shown to be susceptible to adversarial attacks.

3) Maintaining Explainability: If the input distribution is very different from the training data, the AI system will make decisions which may be unexpected or undesired. Preliminary work now exists to try to use AI explainability to counter adversarial attacks and this remains a promising research avenue [19], [20]. The importance of explainability and comprehensibility is set out below.

F. Transparency

The DIN quality pillars introduce the term comprehensibility, which measures the degree to which a stakeholder with defined needs can comprehend the reasons for an AI component's outputs. This is synonymous with explainability. There are wider concerns than explainability relating to transparency. In order to a system to be transparent it is necessary to understand the provenance and quality of input data, including labels, annotations and rules.

Kuwajima & Ishikawa [7] when considering quality models for AI systems, identify transparency as comprising traceability, explainability and communication. This again has a focus on explainability, but includes the ability to communicate the information. Creel [21] identifies transparency in three different forms:

- Functional transparency. Knowledge of the whole operation of the system.
- Structural transparency. Knowledge of the implementation.
- Run transparency. Knowledge of how the system was run in a particular instance.

Functional transparency implies that a human is able to understand the system sufficiently well to accurately predict how an algorithm will behave given particular inputs. This is clearly possible with traditional systems, but with more complex algorithms comprising multiple layers of neural networks and other AI components, it can become nearly impossible.

Structural transparency becomes more important the less it is possible to gain functional transparency. This may be understanding the implementation in code, or could be documentation of the provenance of training data, and statistical analysis done on that data to reduce concerns about accuracy and bias.

Run transparency is the same as post-hoc explainability. Explainability can be ex-ante (before the system has made a decision) or ex-post (after a system has a made a decision). Ex-ante techniques include exploring a dataset to understand and analyse it's various characteristics. This can lead to the identification of class imbalances that heavily influence the systems behaviour [22]. In this context explainability is not a characteristic of the system at all, but a process that is undertaken. Similarly, various mathematical techniques for explainability during the modelling process can be conducted ex-ante, but these merely provide insight into the nature of the system for it's creators.

DeepAI [23] draw a distinction between directly interpretable system that is intrinsically understood by users, and a system that would require an ex-post explanation to understood a specific prediction. It draws a further distinction between a static explanation, and an interactive one that allows users to drill-down and interrogate.

Explainability as a quality characteristic of a system applies only to ex-post explainability. This in itself can take the form

of either internal or external users obtaining an explanation, and communicating it if necessary. Given that, regardless of the method used to obtain an explanation it can be measured in terms of the availability of an explanation (μ), the accuracy of the explanation (α), and the time in which an operator is able to obtain and/or communicate the explanation (T):

$$explainability = f(\mu, \alpha, T)$$

Run transparency is a product quality characteristic that affects the usability of an AI system, and functional and structural transparency most affect the maintainability.

G. Context coverage, context completeness

The operational environments of traditional software systems are typically limited in range and context. The need to define and quantify the impact of the context is recognized in the ISO/IEC 25010 [5] QiU characteristic of *context completeness*:

> [...] degree to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, freedom from risk and satisfaction in all the intended contexts of use, or by the presence of product properties that support use in all the intended contexts of use.

Dynamic AI systems are expected to maintain performance when faced with previously unseen data. This objective is captured by the ISO/IEC 25010 [5] QiU characteristic of *context coverage*:

> [...] degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in both specified contexts of use and in contexts *beyond those initially explicitly identified*. [emphasis added]

Context coverage can be expressed simply as the ability to maintain stability and effectiveness with previously unseen datasets or inputs. Stability can be bought at the cost of statistical bias; by the use of relatively inflexible AI models. Such procedures will lead to under-fitting, where predictions are stable but not accurate enough. By contrast, complex AI models with high-dimensional parameter sets and a low level of statistical bias may be affected by minor features of the training data, which can lead to "over-fitting" and high variance. For supervised learning, a model that is trained to give very accurate results for one dataset will achieve less accurate predictions with new input. There is a direct trade-off between bias and variance [14].

To evaluate a traditional system's context coverage, it would be reasonable to use boundary value analysis or equivalence partitioning [15], perturb data inputs, and measure performance. However this is based on the assumption that traditional programming works using explicit logic and will respond equivalently or proportionally to different inputs. This is not the case with many AI systems that use statistical techniques to arrive at particular outputs. Deep learning in particular has a very large internal feature space which bears little resemblance to the input domain.

The degree to which a system is flexible given an unseen context requires decomposition of the context into a set of input values. For example, to envisage an automated system which links location sensors on a user device with climate control in the user's residence, the inputs could be decomposed as follows:

Input	Data Structure
S. Climate Control Status	boolean
R. Room Temperature	float
T_a. User Minutes Until Arrival	float
T_b. Minutes To Reduce Temp	float
D. User Heading To Room	boolean
D_t. User Heading To Room: Duration	float

We will assume the system has been trained on historical data which exhibit a certain data profile. In this hypothetical example, one can envisage multiple prediction algorithms being used to arrive at a boolean decision that determines whether the system should start cooling the room. Although consistent stability of the procedure in the context of extensive historical data provides some limited quality assurance, QiU context coverage is not guaranteed.

Although the use case is simple, it is clear that a number of the inputs are continuous variables, and as such could have arbitrarily high or low values. It is not possible to measure context completeness within an infinite input domain for an arbitrary AI predictor, even when a reasonable range of intended use is defined, for example with the temperature values restricted to the range $(-100, +100)$, the boolean values 0 to 1, and the arbitrary floats with predefined limits.

Given context completeness is the degree the system maintains performance in an unseen context, it can be measured as the maximum root mean square error (RMSE) for a regression problem (or α and β for a classification problem, see Functional Correctness), that can be obtained by varying any input.

Continuing to use X_i and Y_i to represent ground truth and predicted values respectively, and with n representing the number of observations, RMSE can be defined as follows:

$$RMSE = \sqrt{\frac{\sum (X_i - Y_i)^2}{n}}$$

H. Societal and Ethical Risk Mitigation

AI systems usually intend to treat everyone differently, that is part of their purpose and one reason there is a risk of disparate impact [24]. Issues relating to bias in AI systems sometimes result from propagating existing unfairness in society, known as a societal bias or historical bias [25], or can relate to unfair outcomes resulting exclusively from system implementation. There are metrics that are used to quantify unfairness [26] but the actual nature of unfairness that is to be avoided is derived from the legal, ethical and cultural context of use.

Freedom from risk metrics [5] assess the degree that the quality of the AI system mitigates potential risk to users, organisations and the wider community. Existing quality models focus on health, safety and the environment, however for AI systems it is necessary to consider the wider risks to the rights

and freedoms of members of society. The metrics for health and safety, for instance, tend to relate to reported injuries or hazards. In the context of disparate impact [24] such reporting may not be forthcoming without public investigations such as those by Pro Publica [27].

Fairness is a concept that varies by culture and context, however in the context of AI and automated decision-making, the most prevalent example is that an attribute of person or object unfairly influences the output. Whilst there are many possible metrics for fairness [28], the most generic and flexible way to measure it is through counterfactual fairness [29], as it supports consideration of inputs that are unrelated to fairness in data used for training, however it is unclear how this could be implemented on an existing system, or independent of explainability methods. Given an existing system, there is no difference between the metrics used to measure fairness other than those outlined in the above discussion on bias. Nevertheless, it remains a recommended characteristic for a quality model, because the groups that are measured and acceptable may be different in the context of fairness.

IV. Trustworthiness

The definitions of quality and trustworthiness can be considered different but related, however there are different definitions of trustworthiness. The EU HLEG on AI defines [30] trustworthy AI as

> lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values) and robust (both from a technical and social perspective).

Competing views on this are emerging from the international standards community, the first is that trustworthiness in artificial intelligence is the quality of being dependable and reliable [10]. We can contrast that definition to quality, which is defined by ISO/IEC as *conformance to specified requirements* [5]. A difference between these definitions of trustworthiness and quality, is the need for requirements to be specified by stakeholders, and the verifiability of them. Taking into account the EU view, there is a requirement for systems to deliver against unstated legal, ethical and social requirements as well as technical ones. A second definition is under development within ISO/IEC [31] that defines ICT trustworthiness as a

> demonstrable likelihood that the system performs according to designed behavior under a typical set of conditions as evidenced by its characteristics, such as safety, security, privacy, reliability and resilience.

This definition is based on the NIST Framework for Cyber-Physical Systems [32], and notably includes the work *designed*, which implies the specification of requirements.

Verification is the process of confirmation, through the provision of objective evidence, that specified requirements have been fulfilled. So system owners can make a system trustworthy by specifying verifiable requirements, including consideration for legal, ethical and social issues. Engineers, be they developers or testers, can make a system trustworthy by delivering and verifying requirements, and in theory, stakeholders then trust a system because it is objectively trustworthy. Given the broad scope of the verification, it is very likely that new techniques, business models and certification bodies will spring up in this area.

However, it is not clear that trustworthiness is an independent quality characteristic in its own right, rather it appears to be a superset of a particular set of measurable quality characteristics. Garbuk [33] proposed that a functional characteristics vector could be composed of quality measurements, with appropriate weightings, and that this could be compared to standards for particular AI tasks. These standards would contain measurement methods, minimum quantity of data involved in evaluation, and the minimum observed quality characteristics permitted for a specific task.

V. Proposed Characteristics not included

A. Privacy

Privacy issues are far from unique to AI systems, they can relate to any system that processes personal data. There are numerous pieces of regulation that specifically cover algorithmic decision making [35], which is far more common and complex in the context of AI systems, in comparison to traditional systems. GDPR [36] is the most commonly cited[2], as it provides a right to request a human makes a decision, where a system has made a decision that could have a substantive affect on a data subject.

Beyond the relevance of explainability, there is no obvious unique and novel quality characteristics of privacy relating to AI systems.

B. Collaborability

Some research [7] has suggested that collaborability should be included as an extension to the usability of an AI system. However, metrics are not proposed, and there is limited other literature that covers this topic.

VI. Conclusion

In this paper we have explored numerous aspects of quality for AI systems, their measurements, and their relationship with trustworthiness. This work is not exhaustive, due to the volume and diversity of use cases that AI is being applied to. It is notable that most of the measures proposed, with the exception of controllability and run transparency, are statistical metrics intended to operate across a group of outputs. This speaks to the statistical nature of sub-symbolic AI systems. Whilst most of the measures are not new, they are typically used by system developers and data scientists during the production of a system. These candidate metrics can also be used to evaluate the holistic quality of deployed systems, for which

[2]It is often said that GDPR [36] provides a right to an explanation of how an algorithm reached a particular decision. This is not the case. Whilst it is discussed in the recitals, it is not present in the articles of the regulation as it was removed during the legislative process [37]. Nevertheless, explainable AI is a significant focus for industry as it allows for algorithms that are not well understood to be analysed in order to find metamorphic relationships between groups of system inputs and outputs.

the evaluators may or may not have access to the logic, design, training data or parameters associated with a particular system.

REFERENCES

[1] Y. Zhuang, F. Wu, C. Chen, and Y. Pan, 'Challenges and opportunities: from big data to knowledge in AI 2.0', Frontiers Inf Technol Electronic Eng, vol. 18, no. 1, pp. 3–14, Jan. 2017, doi: 10.1631/FITEE.1601883.

[2] M. Ryan, 'In AI We Trust: Ethics, Artificial Intelligence, and Reliability', Sci Eng Ethics, Jun. 2020, doi: 10.1007/s11948-020-00228-y.

[3] F. Ishikawa and N. Yoshioka, 'How Do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? - Questionnaire Survey', in 2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP), Montreal, QC, Canada, May 2019, doi: 10.1109/CESSER-IP.2019.00009.

[4] 'ISTQB Glossary'. https://glossary.istqb.org/en/search/oracle (accessed Sep. 23, 2020).

[5] ISO 25010. https://iso25000.com/index.php/en/iso-25000-standards/iso-25010 (accessed Aug. 20, 2020).

[6] "ISO - ISO/IEC 25012:2008 - Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model." https://www.iso.org/standard/35736.html (accessed Nov. 02, 2020).

[7] H. Kuwajima and F. Ishikawa, Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems, p. 6

[8] IEC blog - IEC and ISO joint committee on AI adds focus on related data ecosystem. https://blog.iec.ch/2020/05/iec-and-iso-joint-committee-on-ai-adds-focus-on-related-data-ecosystem/ (accessed Aug. 20, 2020).

[9] DIN SPEC 92001-1 - 2019-04 - Beuth.de. https://www.beuth.de/en/technical-rule/din-spec-92001-1/303650673 (accessed Aug. 20, 2020).

[10] "ISO - ISO/IEC TR 24028:2020 - Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence". https://www.iso.org/standard/77608.html.

[11] Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127-131. doi:10.4103/0972-6748.62274

[12] [P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser, 'A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set', arXiv:2003.14263 [cs, stat], Apr. 2020, Accessed: Aug. 20, 2020. [Online]. Available: http://arxiv.org/abs/2003.14263.

[13] K. Yang and J. Stoyanovich, 'Measuring Fairness in Ranked Outputs', in Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago IL USA, Jun. 2017, pp. 1–6, doi: 10.1145/3085504.3085526.

[14] James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, vol. 103. New York, NY: Springer New York, 2013.

[15] S. C. Reid, "An empirical analysis of equivalence partitioning, boundary value analysis and random testing," Proceedings Fourth International Software Metrics Symposium, Albuquerque, NM, USA, 1997, pp. 64-73, doi: 10.1109/METRIC.1997.637166.

[16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[18] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial examples for malware detection. In *European Symposium on Research in Computer Security*, pages 62–79. Springer, 2017.

[19] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, 2018.

[20] Ninghao Liu, Hongxia Yang, and Xia Hu. Adversarial detection with model interpretation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1803–1811, 2018.

[21] K. A. Creel, Transparency in Complex Computational Systems, Philosophy of Science, p. 709729, Apr. 2020, doi: 10.1086/709729.

[22] B. Khaleghi, The How of Explainable AI: Pre-modelling Explainability, Medium, Aug. 15, 2019. https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4 (accessed Aug. 20, 2020).

[23] V. Arya et al., 'One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques', arXiv:1909.03012 [cs, stat], Sep. 2019, Accessed: Aug. 20, 2020. [Online]. Available: http://arxiv.org/abs/1909.03012.

[24] S. Barocas and A. D. Selbst, Big Data's Disparate Impact, SSRN Journal, 2016, doi: 10.2139/ssrn.2477899.

[25] H. Suresh and J. V. Guttag, A Framework for Understanding Unintended Consequences of Machine Learning, arXiv:1901.10002 [cs, stat], Feb. 2020, Accessed: Feb. 22, 2020. [Online]. Available: http://arxiv.org/abs/1901.10002.

[26] S. Verma and J. Rubin, 'Fairness definitions explained', in Proceedings of the International Workshop on Software Fairness - FairWare '18, Gothenburg, Sweden, 2018, pp. 1–7, doi: 10.1145/3194770.3194776.

[27] 'Machine Bias — ProPublica'. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed Sep. 23, 2020).

[28] S. Verma and J. Rubin, 'Fairness definitions explained', in Proceedings of the International Workshop on Software Fairness - FairWare '18, Gothenburg, Sweden, 2018, pp. 1–7, doi: 10.1145/3194770.3194776.

[29] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counter-factual Fairness. In Advances in Neural Information Processing Systems.

[30] ETHICS GUIDELINES FOR TRUSTWORTHY AI, High-Level Expert Group on Artificial Intelligence, EU.

[31] "ISO/IEC WD TS 24462," ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/88/78828.html (accessed Nov. 02, 2020).

[32] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: volume 2, working group reports," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1500-202, Jun. 2017. doi: 10.6028/NIST.SP.1500-202.

[33] A. Kuleshov, "Formalizing AI System Parameters in Standardization of AI", 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI), Nicosia, Cyprus, 2018, pp. 51-54, doi: 10.1109/IC-AIAI.2018.8674446.

[34] M. A. PK, M. R. Sheriff, and D. Chatterjee, 'Measure of quality of finite-dimensional linear systems: A frame-theoretic view', arXiv:1902.04548 [cs, math], Feb. 2019, Accessed: Sep. 30, 2020. [Online]. Available: http://arxiv.org/abs/1902.04548.

[35] A. Chaudhuri, A. L. Smith, A. Gardner, L. Gu, M. B. Salem, and M. Lévesque, 'Regulatory frameworks relating to data privacy and algorithmic decision making in the context of emerging standards on algorithmic bias', p. 6.

[36] European Parliament. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[37] Wachter, Sandra and Mittelstadt, Brent and Floridi, Luciano, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (December 28, 2016). International Data Privacy Law, 2017, Available at SSRN: https://ssrn.com/abstract=2903469 or http://dx.doi.org/10.2139/ssrn.2903469

[38] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, p. 43.