Trustworthy AI Development Guidelines for Human System Interaction

Chathurika S. Wickramasinghe, Daniel L. Marino, Javier Grandio, Milos Manic
Virginia Commonwealth University, Richmond, Virginia
{brahmanacsw,marinodl,grandiogj}@vcu.edu, misko@ieee.org

***Abstract*—Artificial Intelligence (AI) is influencing almost all areas of human life. Even though these AI-based systems frequently provide state-of-the-art performance, humans still hesitate to develop, deploy, and use AI systems. The main reason for this is the lack of trust in AI systems caused by the deficiency of transparency of existing AI systems. As a solution, "Trustworthy AI" research area merged with the goal of defining guidelines and frameworks for improving user trust in AI systems, allowing humans to use them without fear. While trust in AI is an active area of research, very little work exists where the focus is to build human trust to improve the interactions between human and AI systems. In this paper, we provide a concise survey on concepts of trustworthy AI. Further, we present trustworthy AI development guidelines for improving the user trust to enhance the interactions between AI systems and humans, that happen during the AI system life cycle.**

***Index Terms*—Trustworthy AI, Transparency, Explainable AI, Human System Interactions, Human Machine Interactions, AI Life Cycle**

I. Introduction

Artificial Intelligence (AI) nowadays influences all the areas of day to day human activities with the state-of-the-art performance in many areas including health [1, 2], industry [3], natural language processing [1], space exploration [1] and science [4, 5]. Despite their tremendous benefits, many people hesitate to trust AI-based systems due to the black box behaviors, which makes it difficult to get insight into their internal decision making process [6]. In order to build trust between AI systems and humans, it is essential that AI system answer following questions, *Why did you do that?, Why not something else?, When do you succeed?, When do you fail?, When can I trust you?, How do I correct an error?* [7].

To address these trust related issues, the research area of *Trustworthy AI* was introduced recently [8–12]. The goal of Trustworthy AI is to strengthen human trust in AI systems, allowing humans and societies to develop, deploy, and use AI systems without fear and doubt. Many respectful academic and non-academic organizations define trustworthyness as combination of diverse research areas which includes *fairness, robustness, explainability, accountability, verifiability, transperency, and sustainability of AI systems* [8–11, 13–15]. However, as we see, the common goal of these research areas is to improve *human trust* during the *Human System Interactions*.

Human System Interaction (HSI)/ Human AI interactions focuses on design, development, and research on effective interactions between humans and intelligent systems. This includes a wide range of aspects, including intelligent visualizations, human-robotic interaction, virtual/augmented and mixed reality, and visual analytics. These intelligent systems can vary from personal computers, industrial robots, to space stations. Communication between humans and AI systems occurs via various mediums such as user interfaces (GUIs), natural languages, and haptics. During AI system life cycle, three main actors communicate with each other. They are *AI system, System Developers, and System Users*. These actors interact with each other, allowing us to identify three types of HSIs. These three main HSI categories are presented in Figure 1.

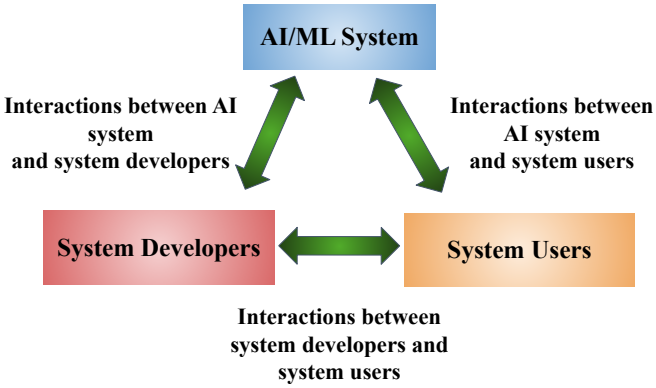

Fig. 1. Human System Interactions during AI System Life Cycle

In this paper, we discuss the *AI Life Cycle (AILC)* and how the three actors communicate with each other in different phases of the AILC. Then we discuss the Trustworthy AI guidelines, which should be implemented during identified human AI interaction categories. Furthermore, we provide a concise survey of existing Trustworthy AI concepts and frameworks.

The rest of the paper is organized as follows. Section II provides a brief survey of Trustworthy AI; Section III discusses AILC and HSI categories, Section IV presents Trustworthy AI guidelines for HSIs. Finally, Section V concludes the paper with a discussion of future research directions.

978-1-7281-7392-4/20/$31.00 ©2020 IEEE

II. Trustworthy AI, Definition, and Principles

This section discusses work related to defining trustworthy AI, its main components, and trustworthy AI principles.

A. Trustworthy AI

Many reputable academic and non-academic organizations (such as Department of Defense, National Science Foundation, IBM, and European Commission) and different domain experts use ethical principles together with formal AI system verification techniques to define trustworthy AI, with the common goal of allowing people and societies to develop, deploy, and use AI systems without fear [8–12]. Ethical principles include the *ethics* of data, algorithms, and practices [12, 16, 17]. AI system *verification techniques* include reliability, resilience, security, and privacy [8, 9]. The rest of this section discusses how individual parties define trustworthy AI and what components are included to verify the trustworthiness of AI systems.

In [16], Floridi et al. pointed out the need for ethical aspects of AI systems. They proposed a trustworthy AI system addressing the ethical impact of three main components: *algorithms, data, and practices*. The ethics of *data* focuses on issues posed by collecting, analyzing, profiling, advertising, and the use of large data sets. The ethics of *algorithms* focus on autonomy and the increasing complexity of ML algorithms and applications. The ethics of *practices* focus on the responsibilities and liabilities of people involved in the AI life cycle and AI systems such as organizations, system users, developers, adopters, and data scientists.

In [12], the *High-Level Expert Group on Artificial Intelligence (HLEGAI)* has taken the first step towards developing a benchmark framework for an ethical AI system. They have comprehensively gathered various ethical principles, offering concrete and practical guidance for multi-disciplinary AI practitioners [17]. The HLEGAI argues that "Striving towards Trustworthy AI concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle". Their proposed framework consist of three components: Ethical principal, seven key requirements with methodologies to implement them, and trustworthy AI assessment list [12].

B. Trustworthy AI Principles

United States, together with the *Organisation for Economic Co-operation and Development (OECD)*, identified five complementary values-based principles for the responsible stewardship of trustworthy AI [11, 13]. The presented OECD trustworthy AI principles are presented in Figure 2. The principles are discussed below,

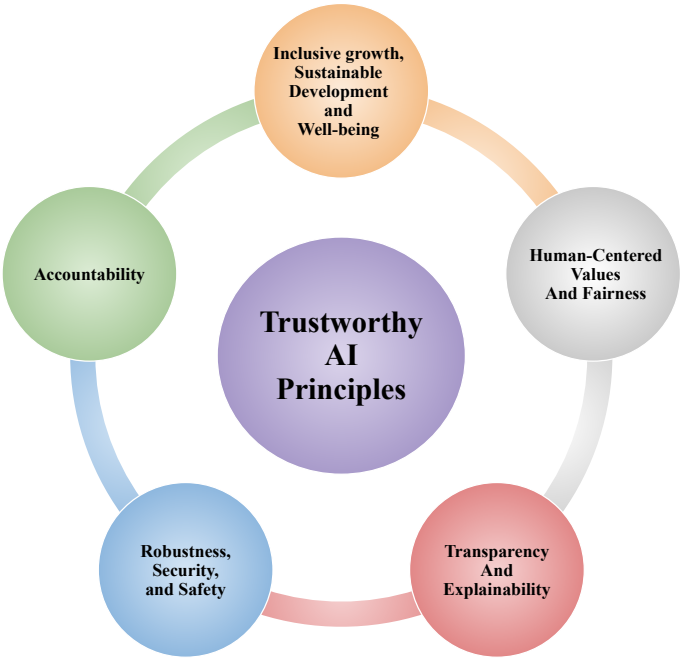

Fig. 2. Trustworthy AI Principles proposed by United States together with OECD [8–11]

1) **Inclusive growth, sustainable development, and well-being:**
Stakeholders should develop, deploy, and use AI systems to benefit humans and the planet. These benefits include augmenting human capabilities, advancing the inclusion of underrepresented populations, and reducing inequalities related to gender and socioeconomic factors [13–15]. Further, AI developments should not harm the natural environment, focusing on sustainable developments and ensuring the well-being of present and future generations [13–15].

2) **Human-centered values and fairness:**
Deployment AI systems should respect the rule of law, human rights, diversity, and democratic values [13–15]. AI actors should implement safeguards and mechanisms to which are appropriate to the context and following with the state-of-the-art. These principles ensure human centered-values such as freedom; privacy and data protection; social justice; labor rights; and equality.

3) **Transparency and explainability:**
The developed AI system should be able to provide a general understanding of the system, which enables those adversely affected by the system to question and challenge its outcomes [13–15]. This includes implementing methods that enable users to understand the outcomes of the AI system plainly and easily.

4) **Robustness, security, and safety:**
The developed AI system should be secure, robust, and safe throughout its entire life cycle. It is essential to identify when these systems fail [18][19], misused, and possible adverse conditions, such that relevant preventive

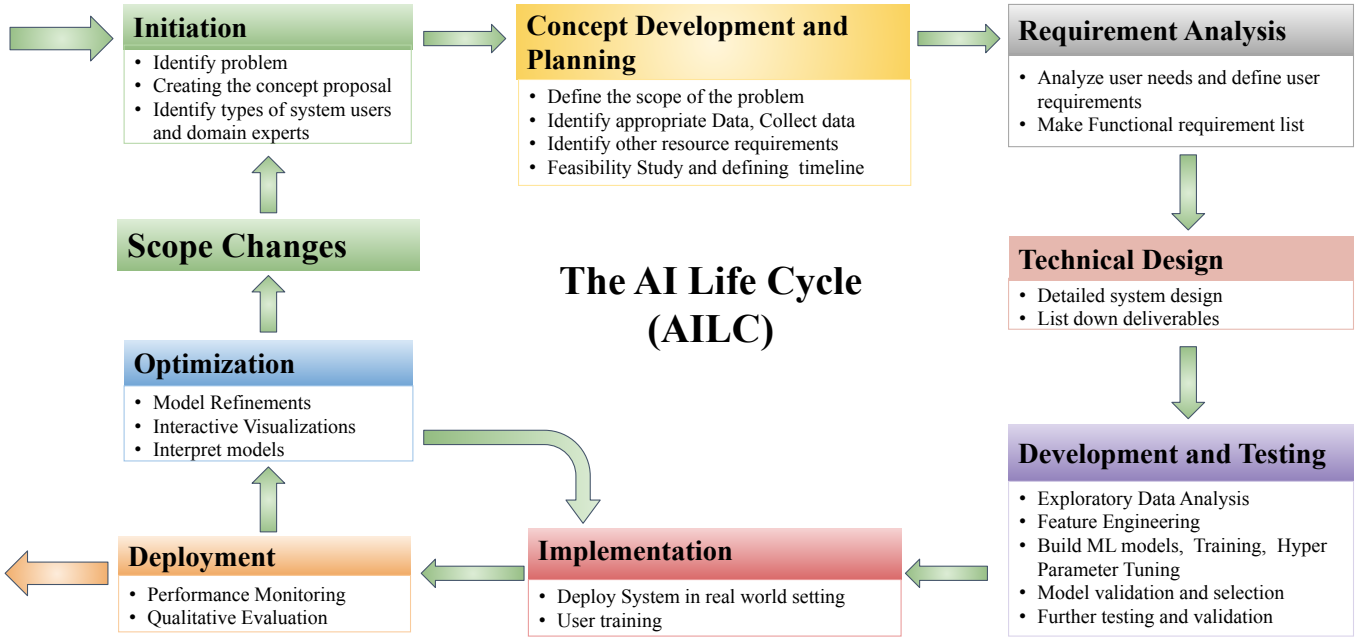

Fig. 3. The AI Life Cycle

mechanisms should be implemented in advance [13–15, 18]. Modeling uncertainty has proven valuable for improving robustness of AI by increasing awareness when there is no enough information to provide accurate estimations [19–22]. AI actors should ensure traceability related to data-sets, processes, and decisions through out the life cycle of AI system. It enables different entities to analyze the system outcomes, respond to inquiries, and ensure the system outcomes are appropriate to the context and consistent with the state-of-the-art. Further, AI actors should apply a systematic risk management approach based on the phase of the life cycle, roles, context, etc. This allows addressing risks related to AI systems, including privacy, digital security, bias, and safety.

5) **Accountability:**
AI actors who involve in designing, development, deployment and use of AI systems should be accountable for the proper functioning of the AI system, respecting the above-discussed principles [13–15].

III. Human System Interactions During AILC

This section discusses the AI Life Cycle (AILC) and three main categories of HSIs during AILC.

Different entities follow different AI system life cycles based on the requirements of a given AI system. However, all of them have common phases and specific tasks which are summerized in Figure 3 [23–27].

In this paper, we categorize HSIs based on the actors involved in the AILC: AI system, Developers, Users. During AILC, three main actors communicate with each other. The identified main actors and relevant examples are presented in Figure 4. As discussed in the Introduction, the interaction between the three main actors can be divided into three categories (Figure 1). The rest of this section discusses the phases in AILC where these interactions take place.

A. *AI and Developers*

In AILC, the phases where interaction between the AI system and System Developers mainly happens are Development and Testing, Implementation, Deployment, Scope Changes, and Optimization phases.

During development and testing, system developers will tightly interact with different software/ML models to explore data, perform feature engineering, model building, model training, model tuning, model testing, model validation, and model selection. During Deployment and Optimization, system developers will perform a quantitative and qualitative analysis of developed AI systems in order to perform model improvements and refinements. During scope changes, developers will identify how the current development can be utilized within the new scope.

B. *AI and Users*

In AILC, the phases where interaction between the AI system and System Users mainly happens are Implementation, Deployment, and Optimization phases.

During implementation, system users will interact with the AI system and evaluate whether the system provides expected deliverables and identify possible refinements to the system outputs. They will provide feedback on how the system performs, whether the system is easy to use, whether the

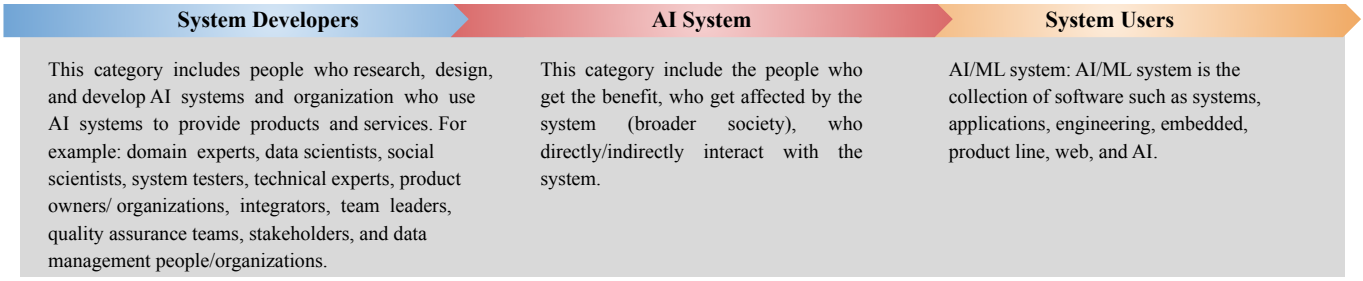

Fig. 4. Actors of Human System Interactions

explanations are provided in an understandable format and whether the explanations are enough.

C. *Developers and Users*

In AILC, the phases where interaction between system developers and system users mainly happens are Initiation, Concept Development and Planning, Implementation, and Optimization phases.

During project initiation, system developers and users interact with each other to identify the problem and roughly-define the scope of the problem. This concept proposal will be further discussed by system developers and users during the concept development phase to define the tight scope of the project. Then system developers agree on data requirements, project timeline, and other resources such as hardware and software.

During Implementation and Optimization phases, the system developers and users will interact with each other, so that system users get trained on how to use the system. In contrast, system developers will identify possible improvements to the system by considering user feedback.

IV. Trustworthy AI Guidelines to Improve the Interactions between Human and AI Systems

This section discusses Trustworthy AI development guidelines which should be implemented to improve the user trust during HSIs. The identified development guidelines are summarized in Figure 5.

A. *AI system and System Developers*

Following guidelines should be implemented to improve the trust of interaction between AI system developers and AI system.

- *Global interpretability:*
  Global interpretability or overall model explanation provides an understanding of the whole logic of the AI system. This gives the entire reasoning process of the system, leading to all the different possible outcomes of the system [6]. Global interpretability is essential for domain experts to analyze whether the developed AI system gives *right outcomes for right reasons*. Further, it allows them to identify what course these AI systems give wrong outputs, allowing them to fix defects and trust the developed system before deployment [28].

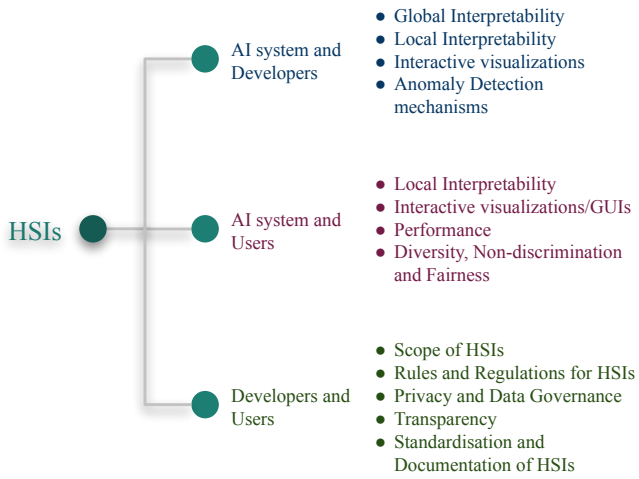

Fig. 5. Trustworthy AI Development Guidelines for HSIs

- *Local Interpretability:*
  Local interpretability or individual prediction explanations are providing why the specific model made a specific decision, why it does not make other possible outcomes [6]. This allows developers to make adversarial samples (changes to input data) and check how the model outcome changes with input data changes. Further, the local explanation can be used to identify harmful interactions such as intentional/unintentional data poisoning, model/infrastructure changes, abusive use of the system, coursing changes to expected AI system behaviors/deliverables. Identifying these prior allows developers to implement safeguard actions prior and to deploy models without fear.

- *Interactive visualizations:*
  It is important to develop interactive visualization on top of the AI system, which allows exploring hidden patterns and model behaviors of the AI system efficiently to the developers[6]. For example, offline system analysis is different from online system analysis because online, the information presented from the system should be short, clear, and easy to understanding. Therefore, based on

the problem context, interactive visualization methods should be implemented. This allows developers to take necessary actions efficiently.

- *Anomaly Detection mechanisms:*
  Developed AI system may encounter new scenarios which it has never seen before. Therefore, it is important to build a mechanism to identify these abnormal/anomalous scenarios, allowing system developers to take necessary actions [29–31]. For example, these anomalies can be due to data drift, or some attacker action. This allows developers to update AI systems and protect them from harmful interactions.

B. *AI system and System Users*

Following guidelines should be implemented to improve the trust of interaction between system users and AI system.

- *Local Interpretability:*
  Local interpretability or individual prediction explanations provide reasons for making a specific decision based on the input from a user. The explanations provided by the AI system should be easy enough to understand by the user. Therefore, the format of the explanations (linguistic, visual, numerical) should be appropriate to the context of the problem and easy to perceive by user. These explanations are essential to build user trust in AI systems as well as to avoid incorrect conclusions about system outcomes. Further, it allows the users to question the decisions made by the system, allowing developers to identify defects and improve the system using user feedback.

- *Interactive visualizations/GUIs:*
  The graphical user interfaces(GUIs) and other visualizations provided by the AI should be user friendly and efficient [32]. Further, AI system should provide wide range of interactive visualizations, covering large audience of users. These visualizations are essential for AI system as it makes AI system easy to to learn and use by users, making users comfortable to use them and trust in them.

- *Performance:*
  In order to build the trust between user and the AI system, it is important to build AI systems with good performance [32]. The performance measures includes but not limited to predictive performance (correct outcomes) and the time take to provide a product or service.

- *Diversity, non-discrimination and fairness:*
  Any interactions between AI system and system users should reflect principles of fairness, which includes avoidance of unfair bias, accessibility and universal design, stakeholder participation [12]. These interactions should not have biases towards certain groups of people (age, gender, abilities, characteristics), which can result in consumer biases and unfair competition between users.

C. *Developers and System Users*

During this category of interaction, the following guidelines should be followed by AI actors.

- *Define the scope of human system interaction during concept development and planning stage of AILC:*
  Scope of HSIs should define during initial stages of the life cycle and refinements should make when necessary. This scope can consist of which entities communicated during what phase, reasons for interactions, what data should be recorded during interactions, who should be aware of what interactions, etc. Further, these documents should be informed to relevant entities prior to the interactions, ensuring the trustworthiness of interactions.

- *Define a set of rules and regulations for HSIs:*
  All entities involve in AILC should define and agree on rules and regulations for possible HSIs. These sets of rules and regulations should be communicated and followed by all entities ensuring the trustworthiness of interactions.

- *Privacy and Data Governance:*
  During these interactions, some data are communicated and recorded. These data may be confidential personal data or any data which should not be exposed to the public. Therefore, the developers and users must agree on privacy and data related regulations defining followings: define what data should be communicated during interactions, what data should be recorded regarding relevant interactions, what kind of data and privacy policies should be implemented, who has access privileges to what portion of data, the reasoning for interactions and data recording, and the lifetime of recorded data.

- *Transparency:*
  All the interactions which happen between developers and system users should be documented in a standard format explaining the reasons for interactions, enabling transparency properties such as traceability and explainability. Further, when developers collect data from system users, it is important to provide a proper understanding of why these data are collected and how they are going to be used.

- *Standardisation and documentation of HSIs:*
  During the life cycle of the AI system, HSIs should be well defined and documented. This allows auditablity, transparency, traceability, and easy refinements when necessary.

V. Discussion, conclusions, and Future Directions

This paper overviews the current research state of *Trustworthy AI*, identifying what principles should be considered when developing, deploying, and use of AI systems. We found that many researchers agree on a set of overlapping Trustworthy AI principles. These include but not limited to *fairness, robustness, explainability, accountability, verifiability, transparency, and sustainability*.

As we see, one main goal of these requirements is to improve human trust during Human System Interactions. Therefore, in this paper, trustworthy AI guidelines were discussed based on different types of human-system interactions that happen during AISLC. Different typed of HSIs were defined based on the three types of actors who interact with each other: *system developers, system users, and AI system*. Defined HSI types were *interactions between system developers and AI system, interactions between system users and AI system, interactions between system developers and system users*.

It has to be noticed that the guidelines for improving human trust during HSIs are context dependant, i.e., depends on the product/services provided by the AI system. For example, for a loan approval system, linguistic explanations of why the request got rejected are more appropriate compared to reject the request without explanation. However, for robot teleportation, a suitable explanation would be haptics or sounds so that in real-time, the operator can understand the actions efficiently. Further, depending on the type of interactions, these guidelines should be different. For example, to trust the developed system, developers need model explanations, whereas system users need individual data explanations.

As mentioned above, the Trustworthy AI research area acts as an umbrella covering diverse research directions. Different academic and non-academic organizations and different domain experts define Trustworthy AI using different sets of overlapping principles/properties. Therefore, we believe that all of these organizations should agree on a *Global framework for trustworthy AI* such that they can build research on top of it.

Most of the recent Trustworthy AI focuses on what principles should be implemented, so that AI systems can be developed, deployed, and use without fear. Machine learning society considers accuracy, precious, recall, and F measures to estimate the goodness of a developed AI system. However, these performance measurements alone are not enough to evaluate the principles of Trustworthy AI. Therefore, we believe that *quantitative and qualitative measures* also should be considered by the research community, focusing on measuring the "Trustworthiness of an AI system". It gives researchers to work on common ground, allowing them to compare AI systems and to verify the trustworthiness formally.

In high-risk areas such as transportation, medical diagnosis, and other mission-critical systems, it is challenging to fully automate AI systems [20]. Because removing humans entirely from the loop can harm the trust of humans in AI systems. In such cases, AI Augmentation is preferred over complete Automation, allowing humans to work side by side with AI systems. Therefore, for high-risk areas, AI Augmentation seems to provide a viable path for building Trustworthy AI [20].

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," en, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] A. Kwasniewska, J. Ruminski, and Szankin, "Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks," *Applied Sciences*, vol. 9, p. 4405, Oct. 2019. DOI: 10.3390/app9204405.

[3] C. S. Wickramasinghe, D. L. Marino, K. Amarasinghe, and M. Manic, "Generalization of deep learning for cyber-physical system security: A survey," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 745–751.

[4] C. S. Wickramasinghe, K. Amarasinghe, and M. Manic, "Deep self-organizing maps for unsupervised image classification," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2019.

[5] K. Czuszynski, J. Ruminski, and A. Kwasniewska, "Gesture recognition with the linear optical sensor and recurrent neural networks," *IEEE Sensors Journal*, vol. 18, pp. 5429 –5438, May 2018. DOI: 10.1109/JSEN.2018.2834968.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.

[7] DARPA, *Explainable artificial intelligence (xai)*. [Online]. Available: https : / / www . darpa . mil / attachments / XAIProgramUpdate.pdf.

[8] U. of Columbia, *Trustworthy ai symposium*. [Online]. Available: https : / / datascience . columbia . edu / trustworthy - ai - symposium.

[9] N. I. of Standard Standards and Technology, *Trustworthy ai: A conversation with nist's chuck romine*. [Online]. Available: https://www.nist.gov/blogs/taking-measure/trustworthy-ai-conversation-nists-chuck-romine.

[10] IBM, *Trusting ai*. [Online]. Available: https://www.research.ibm.com/artificial-intelligence/trusted-ai/.

[11] E. office of the president of the United States, *American artificial intelligence initiative*. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf.

[12] High-Level Expert Group on AI, "Ethics guidelines for trustworthy ai," eng, European Commission, Brussels, Report, Apr. 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[13] T. O. for Economic Co-operation and Development, *Oecd principles on ai*. [Online]. Available: https://www.oecd.org/going-digital/ai/principles/.

[14] OCED, *Recommendation of the council on artificial intelligence*. [Online]. Available: https://www.fsmb.org/siteassets/artificial-intelligence/pdfs/oecd-recommendation-on-ai-en.pdf.

[15] O. T. O. for Economic Co-operation and Development, *Oced ai principals- the role of mps in leveraging the benefits of ai*. [Online]. Available: https://www.oecd.org/parliamentarians/meetings/gpn-meeting-october-2019/Dirk-Pilat-OECD-AI-principles-11-Oct-2019.pdf.

[16] L. Floridi and M. Taddeo, "What is data ethics?" *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, vol. 374, p. 20 160 360, Dec. 2016. DOI: 10.1098/rsta.2016.0360.

[17] N. A. Smuha, "The eu approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, vol. 20, pp. 97–106, 2019.
[18] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 3237–3243.
[19] D. L. Marino and M. Manic, "Modeling and planning under uncertainty using deep neural networks," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 8, pp. 4442–4454, 2019.
[20] D. L. Marino, J. Grandio, C. S. Wickramasinghe, K. Schroeder, K. Bourne, and M. Manic, "Ai augmentation for trustworthy ai:augmented robot teleoperation," in *HSI 2020 - 13th International Conference on Human System Interaction– in press*.
[21] D. Marino, K. Amarasinghe, M. Anderson, N. Yancey, Q. Nguyen, K. Kenney, and M. Manic, "Data driven decision support for reliable biomass feedstock preprocessing," in *2017 Resilience Week (RWS)*, IEEE, 2017, pp. 97–102.
[22] D. L. Marino and M. Manic, "Combining physics-based domain knowledge and machine learning using variational gaussian processes with explicit linear prior," *arXiv preprint arXiv:1906.02160*, 2019.
[23] M. Azure, *The team data science process life cycle*. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle.
[24] J. Pinheiro, *Software development life cycle (sdlc) phases*. [Online]. Available: https://medium.com/@jilvanpinheiro/software-development-life-cycle-sdlc-phases-40d46afbe384.
[25] TrustInsights, *Introduction and the ai/machine learning project lifecycle*. [Online]. Available: https://www.trustinsights.ai/blog/2019/07/5-ways-your-ai-projects-fail-part-1-introduction-and-the-ai-machine-learning-project-lifecycle/.
[26] DataRobot, *Machine learning life cycle*. [Online]. Available: https://www.datarobot.com/wiki/machine-learning-life-cycle/.
[27] DevTeam:Space, *Ai development life cycle: Explained*. [Online]. Available: https://www.devteam.space/blog/ai-development-life-cycle-explained/.
[28] K. Amarasinghe and M. Manic, "Explaining what a neural network has learned: Toward transparent classification," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2019, pp. 1–6.
[29] K. Amarasinghe, K. Kenney, and M. Manic, "Toward explainable deep neural network based anomaly detection," in *2018 11th International Conference on Human System Interaction (HSI)*, 2018, pp. 311–317. DOI: 10.1109/HSI.2018.8430788.
[30] D. L. Marino, C. S. Wickramasinghe, K. Amarasinghe, H. Challa, P. Richardson, A. A. Jillepalli, B. K. Johnson, C. Rieger, and M. Manic, "Cyber and physical anomaly detection in smart-grids," in *2019 Resilience Week (RWS)*, vol. 1, 2019, pp. 187–193.
[31] D. L. Marino, C. S. Wickramasinghe, C. Rieger, and M. Manic, "Data-driven stochastic anomaly detection on smart-grid communications using mixture poisson distributions," in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1, 2019, pp. 5855–5861.
[32] K. Sokol and P. Flach, "Explainability fact sheets," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020. DOI: 10.1145/3351095.3372870. [Online]. Available: http://dx.doi.org/10.1145/3351095.3372870.