Towards a Framework for Characterizing the Behavior of AI-Enabled Cyber-Physical and IoT Systems

Matthew Bundas
Dept. of Computer Science
New Mexico State University
Las Cruces, USA
bundasma@nmsu.edu

Chasity Nadeau
Dept. of Decision & System Sciences
Saint Joseph's University
Philadelphia, USA
cnadeau@sju.edu

Thanh Nguyen
Dept. of Computer Science
New Mexico State University
Las Cruces, USA
thanhnh@nmsu.edu

Jeannine Shantz
Dept. of Decision & System Sciences
Saint Joseph's University
Philadelphia, USA
jeannine.shantz@sju.edu

Marcello Balduccini
Dept. of Decision & System Sciences
Saint Joseph's University
Philadelphia, USA
mbalducc@sju.edu

Tran Cao Son
Dept. of Computer Science
New Mexico State University
Las Cruces, USA
tson@cs.nmsu.edu

***Abstract*— While Artificial Intelligence (AI) and Machine Learning provide a pathway of new and exciting possibilities for AI-Enabled Cyber-Physical and Internet of Things systems, these technology solutions are not without challenges that may hinder adoption. We do not always understand why AI components behave in the way they do, nor can we always predict what they will do under new circumstances. In this paper, we discuss possible approaches for extending the NIST CPS Framework in a way that provides designers, operators and other stakeholders with a shared vocabulary and a collaborative framework allowing them to discuss, identify, express, and verify requirements on the behavior of AI-enabled Cyber-Physical and Internet of Things Systems.**

Keywords—NIST CPS Framework, AI-enabled CPS/IoT, behavior of CPS and IoT

I. INTRODUCTION

Cyber-Physical Systems (CPS) and Internet of Things (IoT) systems frequently pose challenges related to trust, safety, security and assurance in general, especially because many perform safety-critical functions. Formulating requirements on the behavior of Cyber-Physical Systems (CPS) and Internet of Things (IoT) systems is often challenging for the technical and regulatory communities due to the complexity of the issues associated with the design, deployment, and operations of these systems. In response to this challenge, the US National Institute of Standards and Technology (NIST) recently released the *NIST CPS Framework*, which adopts a broad and integrated view of CPS/IoT, providing a shared vocabulary and cross-domain primitives for collaboration on the development of these systems and for the formulation of requirements on their behavior.

The introduction of Artificial Intelligence (AI) enabled components has opened a pathway of new and exciting possibilities for CPS/IoT, but these technology solutions are not without challenges. One critical challenge is in fact the formulation of requirements on the behavior of CPS/IoT that include AI-enabled components. While the NIST CPS Framework provides a useful tool for handling the requirements of the traditional aspects of CPS/IoT, it does not explicitly address the aspects related to AI. Yet, the challenges posed by AI-enabled components are serious to the point that, left unaddressed, they may hinder adoption.

In a nutshell, we do not always understand why AI-enabled components behave in the way they do, nor can we always predict what they will do under new circumstances -- especially when the components leverage "black box" Machine Learning (ML) techniques. These techniques are known to pose challenges well outside of the confines of CPS/IoT. For example, recently the Apple credit card release sparked controversy. David Heinemeier Hansson, Ruby on Rails tech entrepreneur, tweeted alleged gender discrimination in the algorithms used to determine credit limits for the Apple Card. Despite filing joint tax returns, and not disclosing income specifics when applying for the card, Hansson received a credit limit twenty times that of his wife. Ironically, his wife has a better credit score. Apple responded by raising Hansson's wife's credit limit. However, the resolution is a one-off response as Hansson was informed that Apple cannot change the algorithm's decision [1,2]. Apple co-founder Steve Wozniak faced a similar situation, and called on the government to investigate the operation of black box algorithms [3].

While credit limit discrepancies are concerning, and arguably discriminatory, the effects of the unexpected behavior of AI-enabled CPS/IoT in critical infrastructure systems and performing safety-critical functions are potentially catastrophic [4].

In this paper, we discuss possible approaches for extending the NIST CPS Framework in a way that allows stakeholders with a shared vocabulary and a collaborative framework to discuss, identify, express, and verify requirements on the behavior of AI-enabled CPS/IoT.

Next, we provide a brief overview of the NIST CPS Framework. We follow the views from [5,6]. After that, we discuss ways of characterizing the behavior of AI-enabled components and the corresponding framework primitives. Later, we discuss possible approaches for extending the framework and examine their properties by means of a use case. Finally, we draw conclusions and discuss future directions.

II. BACKGROUND: THE NIST CPS FRAMEWORK

The NIST Framework for Cyber-Physical Systems, referred to as "NIST CPS Framework" or simply "Framework" below, comprises a set of concerns and facets related to the system under design or study. This section briefly clarifies the intent and purpose of the framework. The interested reader is directed to SP 1500-201, SP 1500-202 and SP 1500-203, available on the NIST website.

The CPS Framework provides the taxonomy and methodology for designing, building, and assuring CPS/IoT that meet the expectations and concerns of system stakeholders, including engineers, users, and the community that benefits from the system's functions. The Framework comprises a set of concerns about systems, three development facets and a notion of functional decomposition suited to CPS/IoT. The functional decomposition of the Framework breaks a CPS down into functions or sets of functions, as follows: the Business Case, a name and brief description of what the system is or does; the Use Case, a set of scenarios or step-by-step description of ways of using the system and the functions that realize those steps; the Allocation of Function to subsystems or actors – expressed in the terminology of Use Cases; the Physical-Logical Allocation: allocation of given sub-system functions to physical or logical implementation.

The concerns of the Framework are represented in a multi-rooted, tree-like structure (a "forest" in graph theory), where branching corresponds to the decomposition of concerns. We refer to each tree as a *concern tree* of the CPS Framework. The concerns at the roots of this forest are called *aspects*. The Framework comprises nine such aspects, including Timing, Functional, and Trustworthiness. A concern about a given system reflects consensus thinking about method or practice, involved in addressing the concern, and in some cases consensus-based standards describing that method or practice. This method or practice is applied to each function in the functional decomposition of the system and application of a concern to a function results in one or more *properties* (also called *requirements*) to be required of that function in order to address the concern in question. For example, a particular CPS that stores personally identifiable information may pose a confidentiality concern. The Confidentiality concern is a descendant of the Trustworthiness aspect in the corresponding concern tree of the Framework. Thus, the trustworthiness of the CPS is affected. The system's designers may agree that the requirement to use encrypted memory *addresses the Confidentiality concern* and, together with other relevant requirements, addresses its Trustworthiness aspect.

III. CHARACTERIZING THE BEHAVIOR OF AI COMPONENTS

As we discussed, the addition of AI-enabled components to CPS/IoT may complicate the challenges related to understanding their behavior, as well as to formulating and verifying requirements on them. The NIST CPS Framework was not designed explicitly for capturing the behavior of AI-enabled systems.

In our view, the challenge we face bears a link to the general notion of explainability of AI. Specifically, we believe that criteria identified as critical in ensuring the explainability of AI can also be helpful in establishing primitives for characterizing the behavior of AI-enabled CPS/IoT. At this stage, we take inspiration from the Key Performance Indicators (KPIs) and from what George Lawton identified as four ways of making AI more explainable [7]:

1. Understand the data. In addition to having a deep understanding of what the data offers, be sure that training data mirrors the expected data for which the model is developed.
2. Balance explainability, accuracy and risk. Be sure the decisions based on the AI output reflect the company's mission and goals.
3. Focus on the user. Explanations must be appropriate for each stakeholder population. Technical explanations should be reserved for only those groups who understand the language. Understanding is important for promoting end-user trust and adoption.
4. Use Key Performance Indicators (KPI) for AI risk. AI risk may include components: bias, compliance, comprehensiveness, data privacy, explainability, and fairness. Relevant metrics can be generated for each group of stakeholders. [7]

The first item suggests that primitives should be provided for identifying the features of the data used by the system. The second item suggests that certain primitives should be related to notions of explainability, accuracy and risk. The third concern reiterates a foundational notion of the NIST CPS Framework: the vocabulary chosen should be hierarchically organized in such a way that primitives at higher levels of the hierarchy are understandable by all stakeholders regardless of their specific background and interests, while primitives at lower levels of the hierarchy should be focused on particular

classes of experts. The fourth item suggests a potential source for primitives, especially those related to risk, provided by KPIs. KPIs have already been successfully used in a number of domains. One example of KPI adoption related to AI is AI Global's AI Trust Index. This Index is defined as a FICO-like Risk Score for AI. The tool allows companies to define their own best practices and compares AI practices against industry benchmarks [7,8].

Since many companies use KPIs, this is a widely accepted strategy. Joydeep Ghosh Ph.D., chief scientific officer at AI vendor CognitiveScale, claims that companies should first "establish a set of criteria for KPIs for AI risks, including comprehensiveness, data privacy, bias, fairness, explainability and compliance" [7]. In the Apple Card example, the algorithm was not in compliance with New York Law as it resulted in discriminatory treatment of women (or, for that matter, any other protected class of people). This one example hones in on not only compliance, but bias as well.

Leveraging this information provides an indication of useful primitives for an AI-related extension of the CPS Framework. After projecting the above criteria onto the blueprint of the CPS Framework, while at the same time keeping the characteristics of AI-enabled CPS/IoT, the hierarchy of new primitives that we propose to consider is:

- Rationality
    - Compliance
        - Bias
        - Ethics
        - Fairness
    - Comprehensiveness
    - Data privacy
    - Explainability

As the reader may notice, *Rationality* is chosen as the root of the AI-related hierarchy. This is aligned with the view, shared by parts of the AI community, that one of the most salient features of AI is rational behavior. We also find it to be a better choice for the root concept than *AI*, since AI is sometimes viewed as a collection of technology and also because its broad scope overlaps with existing aspects of the CPS Framework rather than being orthogonal to them.

While having a hierarchy of primitives is a step forward, there are several questions to consider before it can be incorporated in the Framework:

1. Should AI and the KPI items be concerns? These risks can include comprehensiveness, bias, fairness, explainability, compliance and data privacy.
2. Is assigning these terms as requirements or properties on existing concerns a more appropriate approach?
3. Do any of these terms fall under the Human aspect?
4. How should any overlap be addressed? For example, as mentioned previously, data privacy is a form of AI risk as identified from KPIs. In addition, privacy is already a concern outlined in the CPS Framework under the Trustworthiness aspect.

IV. TOWARDS A FRAMEWORK FOR AI-ENABLED SYSTEMS

How should the NIST CPS Framework be augmented to support AI-enabled CPS/IoT? We consider three approaches and discuss advantages and disadvantages of each: (1) Preserving the existing concern tree, with no new additions, attempts to capture the elements of the hierarchy through requirements linked to appropriate concerns from the original trees. (2) Disregarding the root of the hierarchy and inserting the other primitives as concerns into an existing suitable concern tree (3) Making *Rationality* an aspect and using the other primitives from the hierarchy as concerns and sub-concerns

In this section, we will evaluate these approaches via an Autonomous Vehicle System (AVS) use case, essentially that of a self-driving car. Portions of concern trees facilitating these approaches are shown in Figures 1-4. In these figures, ovals represent concerns, grey boxes depict properties and, if shown, blue boxes illustrate components. The *sub-concern* relation is represented by the link between two concerns and the relationship *addresses* between property and concern is depicted by a dashed line.

A. AN AVS USE CASE: OBJECT AVOIDANCE.

Use case introduction: The AVS is a complex, AI-enabled CPS, made up of many components allowing the AVS to successfully operate. For simplicity, we limit our discussion to just three core components in the context of this paper. These components are the *Automatic Driving System (ADS), Camera System (CAM)*, and *trainingData*. The *ADS* is responsible for the awareness of the vehicle, as well as making and executing decisions. It makes these decisions from the input it receives, in part from *CAM*, as well as the knowledge and intelligence it has, in part from *trainingData*.

This AVS use case specifically focuses on handling a scenario where an object appears in the direction of travel of a self-driving car. In this situation, the *ADS* is responsible for detecting, classifying, and understanding the object and ultimately, making decisions and applying the best course of action. The *ADS* has been trained from similar scenarios found in the training data, which includes their specific situation, input received, decisions made, actions carried out and their outcome. It uses this prior knowledge to reason and perform its function, and cameras and other data to provide input and awareness in real time.

Requirements and Representation: Here we detail several theoretical requirements of the AVS and how they are represented as properties in the AVS use case:

- *ADS* makes use of several cameras and sensors to provide different angles to help assess the situation -

CAM has property *fiveAngles* indicating five cameras with different angles are operational.

- *ADS* operates at a frequency which allows it to detect issues in real time, and adapt to its environment - *ADS* has property *120Hz* indicating sufficient frequency.
- AVS uses all available data, both previous and current, to assess the situation and have the intelligence to process it - *trainingData* has property *1GB,* considered in this illustrative example to be a sufficient size, and property *exhaustiveData* indicating a sufficient sample of scenarios is present, *ADS* has property *receivingVisual, receivingAudio and receivingTraffic* indicating reception of these data.
- *ADS* determines the most optimal solution to avoid the object or otherwise resolve the situation.
- *ADS* executes the optimal course of actions, continually assessing the situation given changing circumstances.
- *ADS* performs without human-related bias and is sufficiently explainable - *ADS* has property *writingData* indicating data is being written for later reflection and property *peerReviewed* indicating *ADS* has been reviewed and checked for bias and general social compliance, *trainingData* has property balanced indicating fair distribution of scenarios are present.

B. APPROACH. 1: RELYING EXCLUSIVELY ON AI REQUIREMENTS, CURRENT CONCERN TREES

The idea of this approach is to leverage the existing concern trees. AI-related considerations are thus formulated as requirements associated with the most suitable existing concerns. Figure 1 shows the application of this approach for the AVS use case.

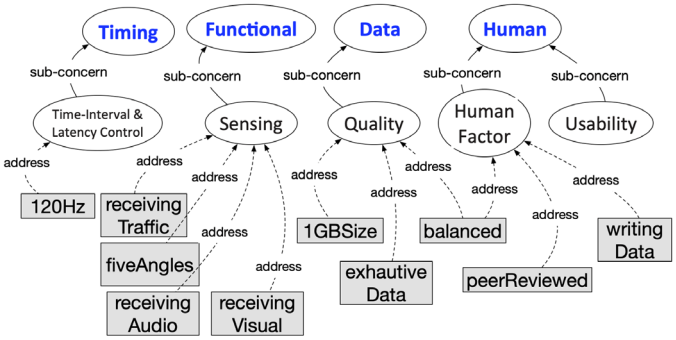

Figure 1 : AVS use case representation using Approach 1, existing concern trees.

In this figure, already existing concerns and AI-related requirements are linked as best as possible. For example, property *peerReviewed* is addressing the *HumanFactors* concern.

This approach has the considerable advantage that it leverages the already existing concerns, which have been carefully vetted by the experts of the Cyber-Physical Systems Public Working Group and are a direct reflection of the way in which experts from different backgrounds traditionally view CPS/IoT. However, this approach also involves a number of challenges. There exist situations where suitable locations and relations for properties involved in the AVS use case cannot be found. For example, in Figure 1, it is hard to see why the *writingData* property is relevant to the *Human* aspect when it is attached to *HumanFactors*. *writingData* is a property which addresses the desire for Explainability in the AVS use case, however *HumanFactors* is a broad enough concern, that this relationship is not precisely represented. A second issue with this approach is that there may be AI-relevant requirements that cannot be captured by any existing concerns. For example, requirements associated with the AVS having intelligence obtained as a result of learning from training data are attached to the *Quality* concern as part of the *Data* concern tree. While this may be the most suitable association in the current CPS Framework, intelligence with regards to AI is a more complicated and nuanced idea than just having "Quality" data as may be indicated in this approach. Lastly, it may be comparatively difficult to both theoretically and computationally reason over AI-relevant requirements and properties in this approach, as they are distributed throughout several concern trees.

C. APPROACH. 2: INTEGRATING AI CONCERNS IN EXISTING CONCERN TREES

The second approach is to create new concerns relevant to AI and integrate them into the existing concern trees to facilitate capturing AI-related considerations. Figure 2 depicts the implementation of the AVS use case based on this approach.

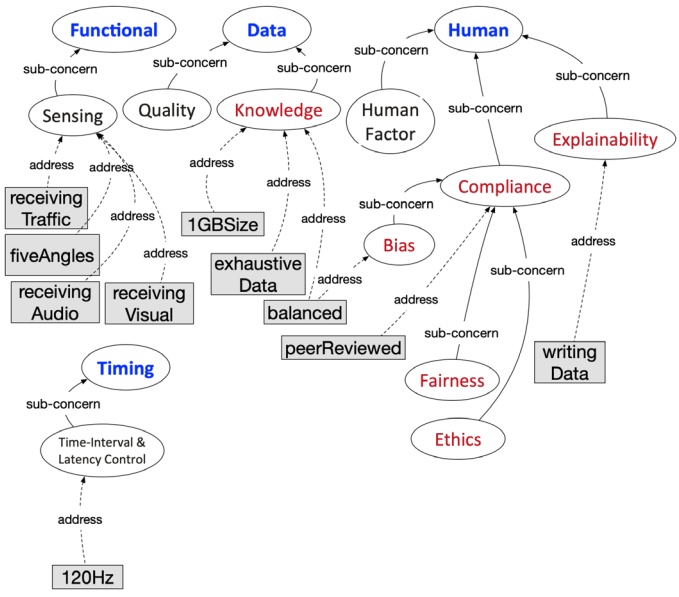

Figure 2 : AVM use case representation using Approach 2, new AI-relevant concerns.

In this figure, in the existing concern trees (i.e. *Timing*, *Functional*, *Data*, etc.), the new AI-related concerns such as *Explainability*, *Compliance*, *Knowledge*, etc. are introduced and attached. For example, *Explainability* and *Compliance* are sub-concerns of the *Human* aspect, *Knowledge* is a sub-concern of *Data* aspect, etc. Note that, as a result of this process, we also refine our initial list of primitives with additional concerns whose role is evidenced by the use case. Figure 2 depicts newly added AI-related concerns with red text, already existing aspects with blue text, and already existing concerns with black text.

In this approach, AI-relevant considerations can be precisely expressed by dedicated concerns while maintaining a clear and clean representation. For example, with the introduction of the *Knowledge* concern, the idea of an AI-enabled CPS having intelligence and (appropriate) knowledge is more meaningfully and effectively represented than in the first approach, where this idea is facilitated using the *Quality* concern in the *Data* concern tree.

One drawback of this approach is that AI-relevant concerns are distributed throughout the concern tree. This makes it difficult to visually pin-point AI-relevant concerns among concerns of other kinds and also has the negative effect of making AI-specific reasoning regarding concern satisfaction potentially more difficult, as several concern trees need to be assessed. For example, in order to verify the satisfaction of *Knowledge* and *Explainability*, the reasoning system needs to consider two concern-trees, *Data* and *Human*.

D. APPROACH. 3: DEVELOPING NEW AI-SPECIFIC CONCERN TREE

In this approach to incorporating AI-relevant concerns, we introduce a new *Rationality* aspect in the CPS Framework that attempts to gather requirements and concerns specifically relevant to AI in an AI-enabled CPS/IoT. The new concern tree associated with this aspect is added to the already existing concern forest. Figure 3 shows the new *Rationality* aspect and its related concerns.

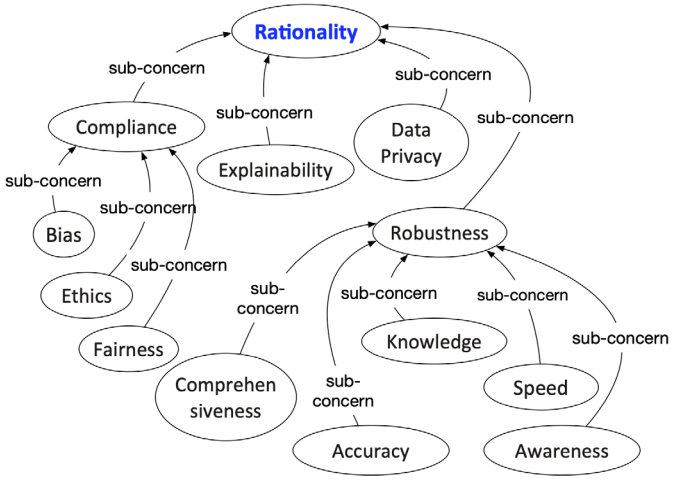

Figure 3 : New AI-related *Rationality* concern-tree

In this figure, the *Rationality* aspect is the root of the AI-related concern tree, and AI-relevant concerns are its sub-concerns. Based on the formalization of the AVS use case, we can capture the relationships between components, properties and AI-related concerns as described in Section IV A. Figure 4 depicts these relationships.

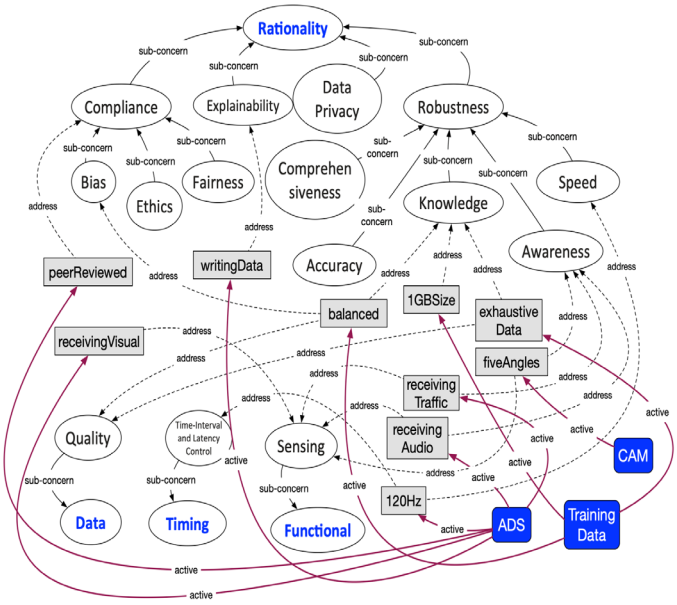

Figure 4 : AVM Use Case representation using Approach 3, with the new *Rationality* aspect.

In this figure, we represent all requirements and properties that are described in the AVS use case, such as the *ADS* controller operating with property *writingData* which addresses the *Explainability* concern.

A benefit of this method of incorporating AI-related considerations into the CPS Framework is that all AI-related concerns are in one place. This helps identify the concerns and reason about the satisfaction of concerns over them easily. A possible drawback is that there are instances of crossover between AI-related concerns in the *Rationality* aspect and other concerns in the forest. This may lead to a potentially less clean and more confusing representation. For example, *120Hz* addresses two concerns, *Speed* in the *Rationality* concern tree, and *Time-Interval and Latency Control* in the *Timing* concern tree. In Approaches 1 and 2, this property only addressed one concern while still capturing the requirement that an AI-enabled CPS operates at a fast speed. To its benefit, in this approach, and similarly to Approach 2, all AI-relevant primitives can be accurately represented, similarly to Approach 2.

V. CONCLUSION

The NIST CPS Framework provides CPS/IoT experts, designers, operators and other stakeholders with a shared set of lenses through which they can collaboratively develop and discuss the essential properties of these systems. However, the CPS Framework lacks the primitives for explicitly capturing the behavior of AI-enabled CPS/IoT and -- most importantly -- for

articulating and verifying its properties. This is a major gap that needs to be filled, as it may hinder reliability and adoption of AI-enabled systems. In this paper, we proposed a first step towards resolving this issue. Specifically, we identified a set of principled AI-related primitives and discussed three possible approaches for incorporating them in the CPS Framework, including their advantages and disadvantages. We hope this will stimulate further work on the refinement of the Framework and lead to a standardized approach and corresponding supporting tools.

Acknowledgements. Portions of this publication and research effort are made possible through the help and support of NIST via cooperative agreement 70NANB19H102.

References

[1] Reuters. "Goldman faces probe after entrepreneur claims gender bias in apple card algorithm." Venturebeat (2019) https://venturebeat.com/2019/11/11/goldman-faces-probe-after-entrepreneur-claims-gender-bias-in-apple-card-algorithm/.

[1] James Vincent. "Apple's credit card is being investigated for discriminating against women." The Verge (2019) https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation.

[2] Shahien Nasiripour & Shridhar Natarajan. "Apple Co-Founder Says Goldman's Apple Card Algorithm Discriminates." Bloomberg.com (2019) https://www.bloomberg.com/news/articles/2019-11-10/apple-co-founder-says-goldman-s-apple-card-algo-discriminates.

[3] P. Laplante, D. Milojicic, S. Serebryakov and D. Bennett, "Artificial Intelligence and Critical Systems: From Hype to Reality," in *Computer,* vol. 53, no. 11, pp. 45-52, Nov. 2020, doi: 10.1109/MC.2020.3006177.

[4] Marcello Balduccini, Edward Griffor, Michael Huth, Claire Vishik, Martin Burns, and David A. Wollman. "Ontology-Based Reasoning about the Trustworthiness of Cyber-Physical Systems" in Living in the Internet of Things: Cybersecurity of the IoT, 2018.

[5] Thanh Hai Nguyen, Tran Cao Son, Matthew Bundas, Marcello Balduccini, Kathleen Campbell Garwood, and Edward Griffor. "Reasoning about Trustworthiness in Cyber-Physical Systems Using Ontology-Based Representation and ASP" in PRIMA-2020: Principles and Practice of Multi-Agents Systems, 2020 (pp. 51-67).

[6] George Lawton. "4 explainable AI techniques for machine learning models." Webpage, <https://searchenterpriseai.techtarget.com/feature/How-to-achieve-explainability-in-AI-models>. Accessed on 4/27/2021.

[7] Manoj Saxena. "Using an AI trust index to unblock stalled machine learning & AI projects." Blog, <https://blog.cognitivescale.com/using-an-ai-trust-index-to-unblock-stalled-machine-learning-ai-projects>. Accessed on 4/27/2021.