Cybersecurity and Innovation Technology Journal, Vol.1, No.1, 2023, pp. 1-15
DOI. 10.52889/citj.v1i1.198

AI Trust Framework and Maturity Model: Improving Security, Ethics and Trust in AI

Submitted 28 February 2023, Revised 26 March 2023, Accepted 8 May 2023

Michael Mylrea[1*], Nikki Robinson[2],
[1,2]Department of Computer Science, University of Miami, Coral Gables, United States
Corresponding Email: *michaelmylrea@gmail.com

Abstract

The following article develops an AI Trust Framework and Maturity Model (AI-TFMM) to improve trust in AI technologies used by Autonomous Human Machine Teams & Systems (A-HMT-S). The framework establishes a methodology to improve quantification of trust in AI technologies. Key areas of exploration include security, privacy, explainability, transparency and other requirements for AI technologies to be ethical in their development and application. A maturity model framework approach to measuring trust is applied to improve gaps in quantifying trust and associated metrics of evaluation. Finding the right balance between performance, governance and ethics also raises several critical questions on AI technology and trust. Research examines methods needed to develop an AI-TFMM and validates it against a popular AI technology (Chat GPT). OpenAI's GPT, which stands for "Generative Pre-training Transformer," is a deep learning language model that can generate human-like text by predicting the next word in a sequence based on a given prompt. ChatGPT is a version of GPT that is tailored for conversation and dialogue, and it has been trained on a dataset of human conversations to generate responses that are coherent and relevant to the context. The article concludes with results and conclusions from testing the AI Trust Framework and Maturity Model (AI-TFMM) applied to AI technology. Based on these findings, this paper highlights gaps that could be filled with future research to improve the accuracy, efficacy, application, and methodology of the AI-TFMM.

Keywords: AI Trust Framework Maturity Model, Cybersecurity, Artificial Intelligence, Machine Learning, ChatGPT

INTRODUCTION

Advances in privacy-preserving machine learning technologies can help improve understanding and trust in with AHMT-S, such as an AI driven digital twin that acts as an industrial immune system (Mylrea et al., 2021a). Previous research demonstrated how digital twins can help improve security and productivity in advanced manufacturing focused on biopharma manufacturing of medical counter measures and vaccines to respond to man-made and naturally occurring biological events (Mylrea et al., 2021b). This article advances that research by filling gaps in Human-AI trust with a focus on improved quantification and metrics of evaluation for an AHMT-S trust framework. For this paper, trust is defined as the contract of assumptions based on humans' perception of and experience with how the system will perform a task or process as designed (Lee & See, 2004). Trust is multi-dimensional and includes a dialectic between humans and machines operating autonomous AI/ML systems. The proper calibration of trust is critical to safe operation of an autonomous system in high assurance environments, such as transportation, health, defense, and other critical infrastructures.

Cybersecurity and Innovation Technology Journal, Vol.1, No.1, 2023, pp. 1-15
DOI. 10.52889/citj.v1i1.198

DISCUSSION

AI Trust Framework Principles

Microsoft has published a set of principles for ethical AI. IBM and Google had very similar models that generally shared the same principles. These principles can be seen in Table 1.

Table 1. Microsoft's principles for ethical AI (Anonymous, nda).

Principles	Explanation
Fairness	AI systems should treat all individuals with fairness, impartiality, and non-discrimination.
Reliability and safety	AI systems should be reliable, safe, and secure,and should operate within the bounds of their intended use.
Privacy and data protection	Microsoft's AI systems should respect the privacy and data protection rights of individuals.
Inclusivity	AI systems should be designed and developed to be inclusive and accessible to people with a diverse range of backgrounds, cultures, languages, abilities, and needs.
Transparency	AI systems should be transparent in their operation and decision-making processes. accountability: Microsoft should be accountable for the design, development, and deployment ofits AI systems, and should ensure that these systems are used responsibly.
Responsible use	AI systems should be used in ways that are consistent with the company's values and principles and should not be used to harm people or society.

AI Trust Framework & Maturity Model (AI-TFMM)

To improve quantification and repeatable measures of trust including security, privacy and ethical controls described above, a maturity model methodology was developed for this research. Maturity models apply weights and/or measurements to specific controls as well as methodology to improve repeatability. This approach is advantageous for measuring performance where adoption of security controls or privacy measures is not a simple yes or no, pass or fail scenario. The proposed approach anchors the AI trust framework above in a methodology that provides a set of ethical principles to help A-HMT-S stakeholders (e.g., Developers, users, organizations, etc.) improve trust in AI. Based on these principles. AI-TFMM provides a common taxonomy and mechanism for AI stakeholders to:

1. Describe the current level or maturity of trust principles,
2. Describe their target state for trust principles,

3. Identify and prioritize opportunities for improvement within the context of a continuous and repeatable process,
4. Assess progress toward the target trust state,
5. Define intended AI uses,
6. Mitigate unethical develop and/or application of ethical,
7. Communicate among internal and external stakeholders' importance of trustworthy AI (Mylrea et al., 2017)

AI-TMM defines four Maturity Indicator Levels (MIL), MIL0 through MIL3, which apply independently to each domain principle. Five aspects of the MILs are important for understanding and applying the AI-TMM:

(1) Maturity Indicator Levels (MIL) Weights and Explanation take a holistic people, process, and technology approach:

Fully Implemented = (3):

Technology: AI trust principles are documented through their lifecycle to be explainable (XAI), repeatable, interpretable, and transparent

People: Someone is assigned/accountable to implementation of these principles through the AI Project lifecycles.

Process: The lifecycle and technology are tested

Largely Implemented = (2)

Technology: AI trust principles are documented through their lifecycle to be explainable (XAI), repeatable, interpretable, and transparent

People: Someone is assigned/accountable to implementation of these principles through the AI Project lifecycles.

Process: The lifecycle and technology are documented but not continuously tested

Partially Implemented = (1)

Technology: AI trust principles are documented through their lifecycle to be explainable (XAI), repeatable, interpretable, and transparent

People: Someone is Not assigned/accountable to implementation of these principles through the AI Project lifecycles.

Process: The lifecycle and technology are documented but not continuously tested

Not Implemented = (0)

Technology: AI trust principles are Not documented through their lifecycle to be explainable (XAI), repeatable, interpretable, and transparent

People: Someone is Not assigned/accountable to implementation of these principles through the AI Project lifecycles.

Process: The lifecycle and technology are Not documented and Not continuously tested

(2) The maturity indicator levels apply independently to each principal domain. AI-TMM users may be operating at different MIL ratings for different domains. For example, an organization could be operating at MIL2 in one domain, MIL3 in another domain, and MIL0 in a third domain.

(3) The MILs are cumulative within each domain; to earn a MIL in each domain, an organization must perform all of the practices in that level and its predecessor level(s). For example, an organization must perform all the domain practices in MIL1 and MIL2 to achieve MIL2 in the domain. Similarly, the organization would have to perform all practices in MIL1, MIL2, and MIL3 to achieve MIL3 (Rosenfeld, 2021).

(4) In applying AI technologies to high assurance use cases, such as defense, patient diagnostics, and other areas where human lives and physical safety can be impacted, a consequence driven approach should be considered. Establishing a target MIL for each domain is an effective strategy for using the AI-TMM to guide ethical AI program improvement. Organizations should become familiar with the practices in the AI-TMM prior to determining target MILs. Gap analysis activities and improvement efforts should then focus on achieving those target levels.

(5) Practice performance and MIL achievement need to align with business objectives and the organization's ethical AI strategy. Striving to achieve the highest MIL in all domains may not be optimal. Companies should evaluate the costs of achieving a specific MIL against potential benefits and document and mitigate any areas where there are gaps in ethical principles. For example, privacy preserving machine learning solutions may run into tradeoffs between efficiency and privacy, transparency, and auditability. The AI-TFMM was developed so that all users can assess and improve ethical AI maturity based on widely accepted principles of trust.

(6) Consistency in measurement of AI algorithms, including transparency and fidelity of data is critical to fostering trust in autonomous human machine teams (Rosenfeld & Richardson, 2019). “Completely, accurately, and clearly quantify the agent’s logic, something that.

Cybersecurity and Innovation Technology Journal, Vol.1, No.1, 2023, pp. 1-15
DOI. 10.52889/citj.v1i1.198

Rosenfeld and Richardson refer to as transparency and Rudin terms fidelity." (Rudin, 2019; Rosenfeld, 2021). Some AI algorithms and technologies, such as optimization models, planning algorithms and semantic reasoning, explainability (XAI) in terms of data inputs and output is less of a challenge. However, various data intensive machine learning models trained on unsupervised neural networks XAI is a challenge. Regardless of the level of complexity, the ethical principles must be considered throughout the AI development and deployment lifecycle to foster trust between human machine teams applying AI technologies.

For example, even with improved transparency of the code, outputs can lack transparency and repeatability if training data is not available for further examination. Thus, the outcome is difficult to explain, communicate and trust; especially AI driven A-HMT-S operating in a high assurance environment. So how could an AI Trust framework help improve trustworthinessand explainability? Take for example, random forest models where you generate a lot of trees randomly – a forest – using different combination of variables that interact. The random forest algorithm tries to find the tree that is most representative for the data. There are several ways to measure the trustworthiness of the algorithm following the ethical AI principles above measured through the AI-TMM. In the context of trust and A-HMT-S, it is critical that a holistic approach is taken to consider the people, process and technology through the AI development, deployment, and management lifecycle. AI lifecycles can vary depending on the model and use case; however, the five general stages are:

1. Problem Scoping,
2. Data Acquisition,
3. Data Exploration,
4. Modelling and
5. Evaluation

These steps can be broken down into several sub tasks: i. Define Project Objectives, ii. Acquire and Explore Data, iii. Model Data, iv. Interpret & communicate, v. Implement, Document & Maintain as highlighted in Figure 1.

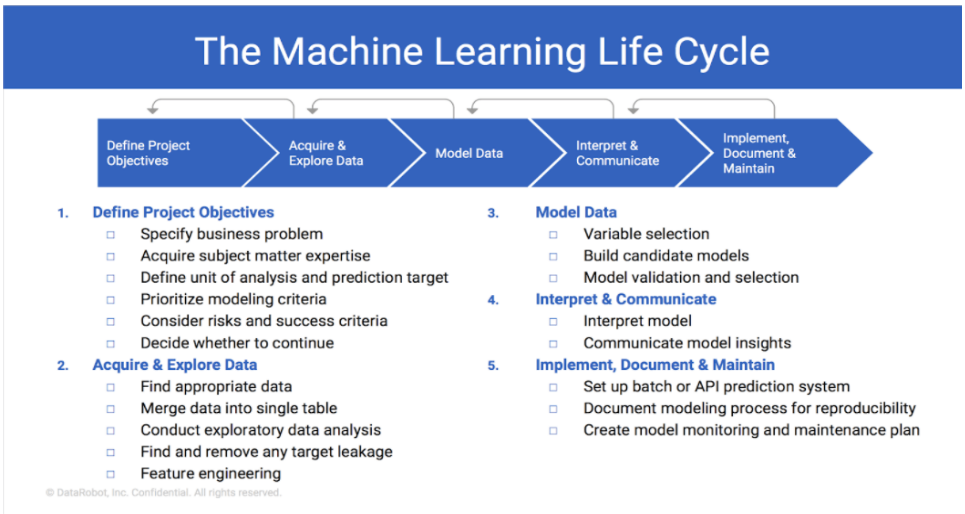

Figure 1. Highlights Machine Learning Lifecycle (DataRobot) (Anonimous (ndb).

Continuing with the example of random forest model, to validate the trustworthiness from a holistic perspective (technology, people, and process). A couple of important areas of focus would include, but not be limited to understanding the accuracy of model. How is it being applied? What data sets are being used? Do we understand the code? There are number of ways to test the accuracy of the technology, but in considering trust is social contract between human and machines and underlying assumptions – this can't be done without people and process consideration. So, we start here with the technology and examining the code. The following is Python script for anomaly detection using a random forest model:

```
import numpy as np

from sklearn.ensemble import RandomForestClassifier

# Load the training data
X_train = np.load("X_train.npy")
y_train = np.load("y_train.npy")

# Load the test data
X_test = np.load("X_test.npy")
y_test = np.load("y_test.npy")

# Create the random forest model
model = RandomForestClassifier()

# Train the model on the training data
model.fit(X_train, y_train)

# Use the model to predict labels for the test data
y_pred = model.predict(X_test)

# Calculate the accuracy of the model
accuracy = np.mean(y_pred == y_test)
print("Accuracy: ", accuracy)
# Identify anomalies in the test data
```

```
anomalies = np.where(y_pred == -1)[0]
print("Anomalies: ", anomalies)
```

This script assumes that you have already loaded the training and test data into NumPy arrays X_train, y_train, X_test, and y_test. While beyond the scope of this study, future research should apply this script and test the AI-TMM against the trustworthiness principles of A-HMT-S via the random forest model on the training data and uses the model to predict labels for the test data. It then calculates the accuracy of the model and identifies the indices of any anomalies in the test data. Additional testing is needed to determine the level of trust as it relates to the ethical principles that underpin the AI-TMM. Let's examine the AI trust maturity level of the technology. The following questions should be considered in this evaluation:

Technology: Are AI trust principles documented through their lifecycle to be explainable (XAI), repeatable, interpretable, and transparent?

- If yes, then maturity indicator level = MIL 3 or fully implemented
- If principles are in place, documented and managed, but they are not tested = MIL 2 or largely implemented
- If principles are in place and documented but not managed or tested = MIL 1 or partially implemented

People: Is someone assigned/accountable to implementation of these principles through the AI Project lifecycles?

- If yes, then maturity indicator level = MIL 3 or fully implemented
- If principles are in place, documented and managed, but they are not tested = MIL 2 or largely implemented
- If principles are in place and documented but not managed or tested = MIL 1 or partially implemented

Process: The lifecycle and technology are continuously tested?

- If yes, then maturity indicator level = MIL 3 or fully implemented
- If principles are in place, documented and managed, but they are not tested = MIL 2 or largely implemented
- If principles are in place and documented but not managed or tested = MIL 1 or partially implemented

The accuracy of the model will depend on the quality of the training data and the chosen model parameters. Thus, ethical people and processes consideration are just as important as the model. In the python script above, the model is a random forest classifier with default parameters. In general, random forests are quite robust and can produce accurate results on a wide range of tasks. However, the accuracy of the model will depend on the characteristics of the dataset and the use case they are being applied to. It is important to evaluate the model's performance on a test set or using cross-validation to get a good estimate of its accuracy.

The script provided calculates the accuracy on the test set using the line accuracy = np.mean(y_pred == y_test). This compares the predicted labels (y_pred) to the true labels (y_test) and calculates the fraction of predictions that are correct. It is also important to consider other evaluation metrics in addition to accuracy, such as precision, recall, and AUC (area under the curve). These metrics can provide a more complete picture of the model's performance, especially when dealing with imbalanced datasets.

Reproducibility of results is another critical principle for trust in AI by measuring:

1. The ability to reproduce outcomes,
2. Reconstruct how features were developed and selected,
3. Understand the interaction of the features,

For example, can you replicate a local optimum of the model used for a single decision of the model to estimate the global optimum? In comparing the local and global optimum can you determine what features have a strong impact on a decision, and which are the key features overall? Transparency, reproducibility and explainability on how AI technology is trained, inputs and outputs is all part of improving trust in AI driven A-HMT-S. Figure 2 highlights various inspection methods to improve transparency, ethics, and trust in AI. Note data quality is a critical dimension to all the AI models and inspection methods. Thus, the data acquisition lifecycle, including people and process, along with technology is imperative. In addition to reproducibility, audibility, transparency and explainability are also critical factors in assessing trust.

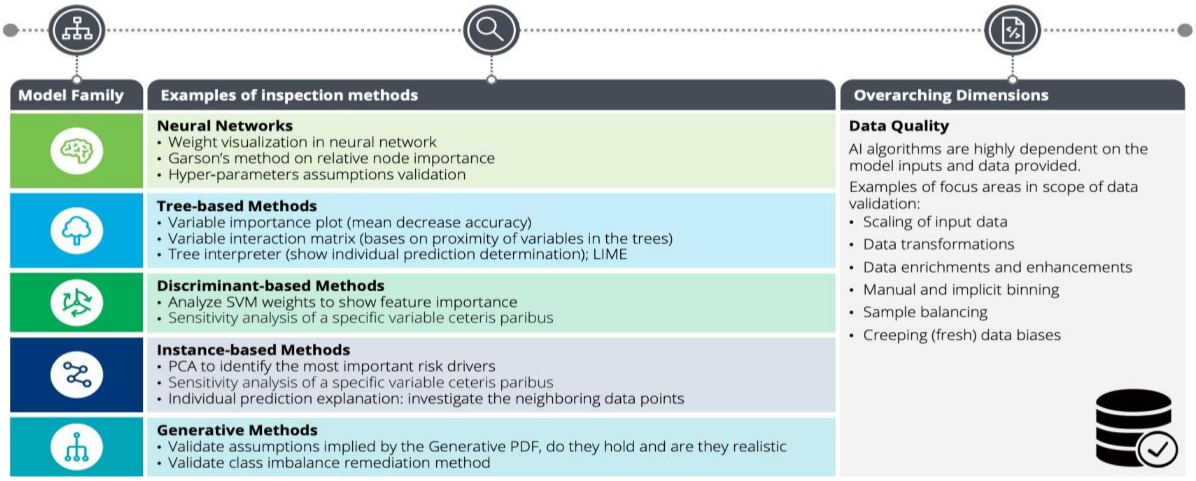

Figure 2. Highlights AI technology inspection methods to improve explainability, transparency and trust (Deloitte, 2021).

DOI. 10.52889/citj.v1i1.198

Explainable Artificial Intelligence (XAI) is a field of Artificial Intelligence (AI) that promotes a set of tools, techniques, and algorithms that can generate high-quality interpretable, intuitive, human-understandable explanations of AI decisions. ML, called Deep Learning (DL), where thousands (even billions) of neuronal parameters are trained to generalize on carrying out a particular task (O'donovan et al. 2015). Developing intuitive, trustworthy, explanations of how and why our AI models arrive at decisions is critical to improving trust in A-HMT-S. Trust requires AI models that are easy to interpret, visualize and use. When you have numerous parameters in Deep Neural Networks (DNNs) it adds complexity that is hard to interpret. Algorithms and models can learn and misinterpret representations from the data differently than humans creating bias, error, and lack of trust. Improving AI-TMM people and process maturity can help avoid cognitive bias. Bias can have a significant impact on the fairness and trustworthiness of the AI technologies application, training data and process and results. Highlights how cognitive bias in data models and inputs reduces trust, fairness, and ethics of outputs of AI technologies can be seen in Figure 3.

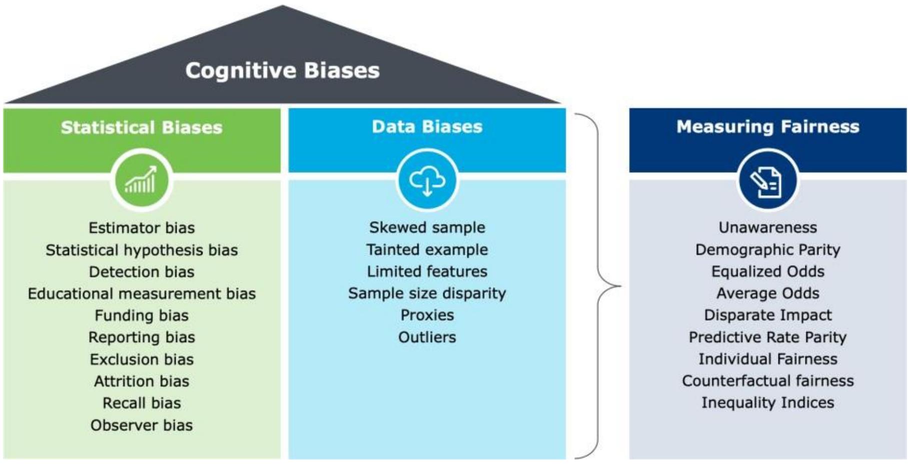

Figure 3. Highlights how cognitive bias in data models and inputs reduces trust, fairness, and ethics of outputs of AI technologies (Deloitte, 2021).

To improve trust, transparency and XAI for a more ethical AI we must ensure that data used to train and validate AI models is free of cognitive bias and representative of a diverse population. Even if the AI model is well-designed, if the input data is biased, the model will replicate that bias in its outputs. If the data sets are available, then they can be audited to detect bias in data sets. Consider for example ChatGPT and the lack of transparency and auditability on its outputs: What data sets were used? How were the prepped, classified and aggregated for training? What safeguards were implemented to mitigate cognitive bias? Was private, proprietary or any sensitive data used during the training process? The proposed AI-TMM trust framework could potentially be used to audit and/or design AI technologies, such as ChatGPT for biases in the data before it is used to train a model. To improve trustworthiness, it is critical to test for biases using various scenarios. However, it can be difficult to detect bias in a model until it is in use, as it is often only through observing the model's outputs that biases can be identified. It is therefore important to consider the ethical implications of AI models and to work towards developing more explainable and unbiased AI technologies to improve trust between AI to improve autonomy in human machine teams (Roovers, 2019).

Social scientists have been challenged to reproduce their findings (Hatchwell, 2017) via independent data models such as Shannon information theory (Shannon, 1948). Confirmation bias, gaps in transparency and fidelity of data used in training and developing AI algorithms also creates a lack of trust in the assumptions that underpin these theories to enable predictability of outcomes. This limits their value and applicability to generalize relates assumptions needed to solidify a contract of assumptions to enable predictability between autonomous human-machine systems. To validate this theory this study concludes by testing the AI trust frameworks controls against ChatGPT - the popular chatbot from OpenAI, which can be considered an A-HMT-S. There is an increasing number of examples from drones to self-driving cars, precision medicine and diagnostic tools, robots, and virtual training platforms that are being trusted with critical decisions. Some scholars argue that no critical decisions should be trusted to black box machine learning models (Rudin, 2019). Applying the AI trust framework and lenses such as XAI is critical to understanding if A-HMT-S can be trusted? How much can they be trusted? When should humans in the loop get the final say versus a chatbot? What legal jurisdiction is the chatbot? And who should be held accountable as its outputs are incorporated to critical decision making?

Validation of AI-TFMM Examination of ChatGPT

To validate and verify the efficacy of the AI-TMM, the examination tests key metrics of evaluation against Open AI’s ChatGPT. OpenAI's GPT, which stands for "Generative Pre-training Transformer," is a deep learning language model that can generate human-like text by predicting the next word in a sequence based on a given prompt. ChatGPT is a version of GPT that is tailored for conversation and dialogue, and it has been trained on a dataset of human conversations to generate responses that are coherent and relevant to the context. ChatGPT is one of many language models developed by OpenAI and is intended for use in research and development in the field of natural language processing.

ChatGPT is a machine learning tool that can generate human-like text based on a given prompt. Some potential uses for ChatGPT include customer service and support, education, training, and personal assistants. However, there are also concerns about the potential impacts of this technology, such as job loss and reduced human interaction. ChatGPT was trained on a large dataset of text from the internet, and it was tested by having people rate its responses. OpenAI, the company behind ChatGPT, is focused on researching and developing artificial intelligence in a way that is beneficial to humanity.

Unlike some AI/ML language models, ChatGPT does not use GANs. Instead, Chat GPT uses a transformer architecture, which is a type of neural network that is particularly well-suited to processing sequential data such as text. GANs, on the other hand, are a type of machine learning model that is used to generate synthetic data that is like a given training dataset.GANs consist of two neural networks: a generator network that produces synthetic data, and a discriminator network that tries to distinguish the synthetic data from real data. The generator and discriminator networks are trained together in a adversarial process, with the generator trying to produce synthetic data that is indistinguishable from the real data, and the discriminator trying to correctly identify which data is real and which is synthetic. GANs have been used to generate a wide range of synthetic data, including images, audio, and text. However, they are not directly related to the GPT language model.

Question to ChatGPT: Is Open AI's ChatGPT trustworthy?

Response: “As an AI language model trained by OpenAI, I am not able to evaluate the trustworthiness of other systems or technologies. However, I can tell you that OpenAI is a well-respected research organization in the field of artificial intelligence, and they are committed to developing and promoting responsible and ethical use of AI. I would recommend that you carefully consider your needs and requirements and do your own research and due diligence before making any decisions about whether to use a particular technology.” (Anonimous, ndc).

A full scope assessment of each of the 5 ethical AI principles is beyond the scope of the study. Thus, an examination of XAI is prioritized for this research.

AI-TMM Assessment Explanation:

Technology: Are AI trust principles documented through their lifecycle to be explainable (XAI), repeatable, interpretable, and transparent?

ChaptGPT does provide documents on how their model is trained. This model is documented and managed. Thus, one can argue that Principles are in place, documented and managed and tested = MIL 3. The process is shown in the figure below. To create a reward model for reinforcement learning, ChatGPT collected comparison data, which consisted of two or more model responses ranked by quality. To collect this data, ChatGPT developers took conversations that AI trainers had with the chatbot, suggesting that this process is documented and managed (requirement for MIL 2). Finally, in the realization of MIL 3 the model is

Cybersecurity and Innovation Technology Journal, Vol.1, No.1, 2023, pp. 1-15
DOI. 10.52889/citj.v1i1.198
tested.

ChatGPT documentation below highlights how developers "randomly selected a model-writtenmessage, sampled several alternative completions, and had AI trainers rank them. Using these reward models, [developers] can fine-tune the model using Proximal Policy Optimization. Documentation also highlights that the testing process is continuous, consisting of "several iterations of this process." (Anonimous, ndd). Highlights Open AI ChatGPT documentation for training methods can be seen in Figure 4.

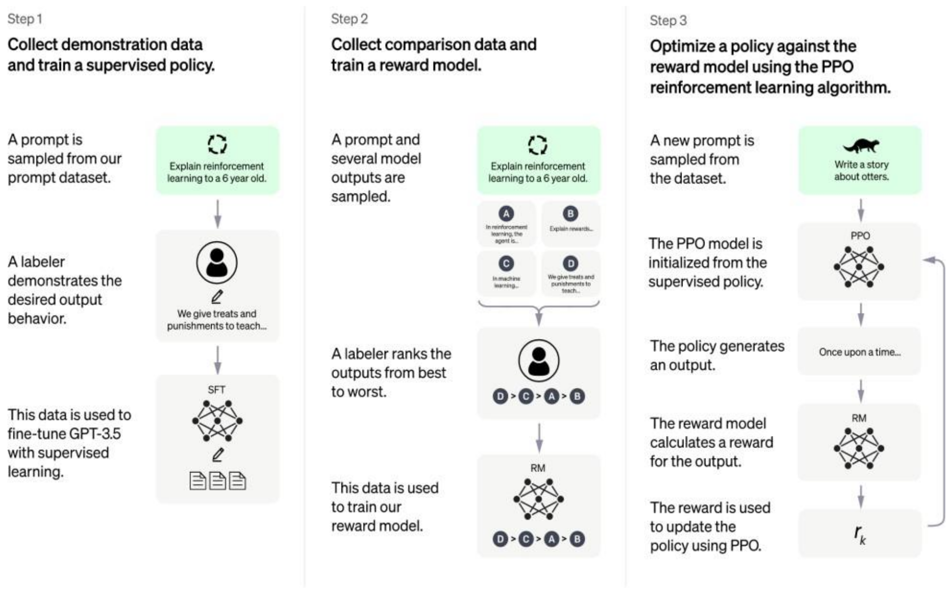

Figure 4. Highlights Open AI ChatGPT documentation for training methods (Anonimous, nde)

However, one could question the maturity indicator level highlighting the need for improved transparency of the code and data used for training. Indeed, the accuracy of the model will depend on the quality of the training data and the chosen model parameters.

People: Is someone assigned/accountable to implementation of these principles through the AI Project lifecycles?

Documentation suggests that OpenAI assigned developers to implementation of ethical AI trustworthy principles through the AI Project lifecycles or MIL 3. However, there are limitations in auditability that could question the veracity of this weight. For this reason, ChatGPT outputs often include plausible sounding but incorrect or nonsensical answers. Quantifying the level of trustworthiness is challenging as ChatGPT documentation suggests: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly

Process: Is the lifecycle and technology continuously tested?

Similar limitations exist for measuring the AI-TMM for process. The documentation suggests MIL 3 in that AI ethical principles are in place, documented, managed, and tested (MIL 3. However, ChatGPT documentation also notes an inherent gap in the trustworthiness ofthe model "Supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows." AI-TMM assessment of ChatGPT transparency as assessed via the lens of AI ethical principles incorporated to the people, process and technology lifecycle of the AI model and application can be seen in Table 2.

Table 2. AI-TMM assessment of ChatGPT transparency as assessed via the lens of AI ethical principles incorporated to the people, process and technology lifecycle of the AI model and application.

Principles	Definition	AI-TMM Assessment Score
Fairness	AI systems should treat all individuals with fairness, impartiality, and non-discrimination.	Out of scope, but an important and timely target for future research.
Reliability and safety	AI systems should be reliable, safe, and secure, and should operate within the bounds of their intended use.	Out of scope, but an important and timely target for future research.
Privacy and data protection	AI systems should respect the privacy and data protection rights of individuals.	Out of scope, but an important and timely target for future research.
Inclusivity	AI systems should be designed and developed to be inclusive and accessible to people with a diverse range of backgrounds, cultures, languages, abilities, and needs.	Out of scope, but an important and timely target for future research.
Transparency	AI systems should be transparent in their operation and decision-making processes. accountability: Microsoft should be accountable for the design, development, and deployment of its AI systems, and should ensure that these systems are used responsibly.	People - MIL 2 Process - MIL 2 Technology- MIL 2 ***Questions raised on methodology limitations in applying to certain AI algorithms
Responsible use	systems should be used in ways that are consistent with the company's values and principles and should not be used to harm people or society.	Out of scope, but an important and timely target for future research.

CONCLUSION

Data is the new gold of the Digital Age. Big data is critical for Artificial Intelligence (AI) and Machine Learning (ML) driven predictive models. Data collection and use in training machine learning algorithms has created new security, privacy, and ethical challenges. Data exfiltration and manipulation can cause significant financial, reputational, and even physical harm. This requires advanced in explainability of how AI/ML algorithms are arriving at their conclusions. Improved explainability will help improve trust only if it can be measured and understanding is intuitive. This requires humans in the loop to understand how the algorithms are being trained and arrive at their conclusions. While solving for explainable AI or XAI, however, this in turn creates additional privacy challenges. How can an autonomous system be trained on sensitive data sets that are also explainable if a human in the loop can’t consume the data due to privacy concerns?

These improved metrics examine how risk is weighed and decisions are made. This understanding is key issues related to interdependence and dependence when faced with uncertainty, conflict, and competition. The contextual dynamics of measuring trust as social construct and trust anchor in an era of uncertainty and change is critical to establishing metrics that can be used to optimize performance and predictability. Failed predictions from intelligence agencies on Russia’s goals for Ukraine to economic performance to impacts of Covid-19 are just some of the potential global implications for misunderstanding entropy production. This article highlights the importance of PPML in increasing the trustworthiness and subsequent autonomy of critical systems with humans in the loop. While we are quick to leverage solutions such as Open AI’s Chat GPT to expedite data research and data collection, we are reverent for humans in the loop for critical decision making; especially when our lives or limbs are at risk. We are quick to trust proven humans even forgive where error or risk is explainable. The same luxury is not granted to machines or AI/ML algorithms. In examinationof a popular AI ChatGPT suggests there several trust gaps that need to be filled.

SUGGESTIONS

Additional studies should be conducted to better understand how to make generative AIsolutions - like ChatGPT – more trustworthy. These studies should leverage the AI Trust Maturity model developed for this paper to evaluate and improve ethical AI principles in other generative AI solutions. These studies are critical to improving our understanding of these models’ trustworthiness, transparency and explainability. A critical aspect of this future research should include improved fidelity and transparency of training data, which is one of thebiggest gaps found in determining the trustworthiness of the model. A holistic people, processand technology approach bolsters the contextual understanding of its application, but also introduces challenges for repeatability for different use cases.

REFERENCES

Anonimous (nda). Microsoft's principles for ethical AI. Accessed on January 3, 2023 at INSERT

Anonimous (ndb). Data Robot. Machine Learning Life Cycle. Accessed on January 4, 2023 at https://www.datarobot.com/wiki/machine-learning-life-cycle/

Anonimous (ndc). "ChatGPT" (December 15 Model). OpenAI. Accessed on January 2, 2023

Anonimous (ndd). ChatGPT AI development and training methods. Accessed on January 2, 2023, at https://openai.com/blog/chatgpt/

Anonimous (nde). ChatGPT Documentation. Accessed on Jan 2, 2023 at https://openai.com/blog/chatgpt/

Deloitte, A. B. (2021). 2021 Transparency Report.

Hatchwell, B. J. (2017). Replication in behavioural ecology: a comment on Ihle et al. Behavioral Ecology, 28(2), 360-360.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

Mylrea, M., Gourisetti, S. N. G., & Nicholls, A. (2017, November). An introduction to buildings cybersecurity framework. In 2017 IEEE symposium series on computational intelligence (SSCI) (pp. 1-7). IEEE.

O'donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. T. (2015). Big data in manufacturing: a systematic mapping study. Journal of Big Data, 2, 1-22.

Roovers, R. (2019). Transparency and responsibility in artificial intelligence. A call for explainable AI. Accessed on December 25, 2022 at https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf

Rosenfeld, A. (2021). Better metrics for evaluating explainable artificial intelligence. In Proceedings of the 20th international conference on autonomous agents and multiagent systems (pp. 45-50).

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems, 33, 673-705.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.