Paper ET02

A Holistic Approach to Trustworthy Artificial Intelligence (AI)

Qingying (Ally) Lu, IBM, Washington DC, USA
Tara Chavda, IBM, Washington DC, USA
Kathryn Matto, IBM, Washington DC, USA

ABSTRACT

Over the past decade, Artificial Intelligence (AI) has advanced significantly to be able to perform many tasks previously done by humans. The key advancement in the AI technology is that algorithms can recognize patterns in data and build models to recommend decisions without requiring explicit programming. Since the AI models learn behavior from the training data, any patterns (e.g., biases) in the training data will be faithfully reproduced in the model behavior. Due to the statistical nature of the models, they are also likely to make errors due to natural scatter in the data and uncertainties in the chosen models. In addition, AI systems often operate as “black boxes” and may produce unfair outcomes and cause other ethical concerns. Therefore, fostering trustworthy AI practices is not only a responsible and ethical choice but also a strategic imperative for organizations looking to build trust, maintain compliance, and thrive in an increasingly AI-driven landscape.

In this paper, we will present a comprehensive AI governance framework for overseeing the development, deployment, and usage of AI to foster trustworthy AI practices within an organization. This comprehensive framework necessitates governance not only at the organizational level but at the AI model operational level. Establishing AI governance at the organizational level, multi-disciplinary teams are essential for a holistic approach to this socio-technical challenge. At the operational level, we advocate a multi-faceted approach to assessing AI models, encompassing the following five key dimensions: (i) Competence, (ii) Fairness, (iii) Robustness, (iv) Openness, and (v) Privacy.

INTRODUCTION

Over the past decade, Artificial Intelligence (AI) has advanced significantly to be able to perform many tasks previously done by humans. The key advancement in the AI technology is that algorithms can recognize patterns in data and build models to recommend decisions without requiring explicit programming. Since the AI models learn behavior from the training data, any patterns (e.g., biases) in the training data will be faithfully reproduced in the model behavior. Due to the statistical nature of the models, they are also likely to make errors due to natural scatter in the data and uncertainties in the chosen models. In addition, AI systems often operate as “black boxes” and may produce unfair outcomes and cause other ethical concerns. Therefore, fostering trustworthy AI practices is not only a responsible and ethical choice but also a strategic imperative for organizations looking to build trust, maintain compliance, and thrive in an increasingly AI-driven landscape. The advent of Large Language Models (LLMs) accentuates the necessity of ensuring trustworthiness in AI, given their heightened impact. Recent developments, such as the White House AI Executive Order, underscore the importance of safe, secure, and trustworthy AI development.

The pursuit of AI Trustworthiness has evolved into a multidisciplinary field, emphasizing the need to prioritize user requirements and embrace an ethical approach that encompasses people, processes, and technology. Fostering AI Trustworthiness within an organization requires the implementation of a robust AI governance framework. This framework aids in responsibly designing, developing, and utilizing AI technologies, ensuring the integration of trustworthiness considerations throughout the AI life cycle - from strategy and planning to development, operations, and monitoring.

A COMPREHENSIVE AI GOVERNANCE FRAMEWORK

“AI Governance Strategy aligns with intended business outcomes and technology. It encodes policies into business rules and transparent reporting mechanisms to establish intentional clarity [1]” IBM advocates for a comprehensive AI governance framework for overseeing the development, deployment, and usage of AI to foster responsible and ethical AI practices within an organization. IBM has invested in a suite of efforts to establish an AI Governance framework for developing principles and internal processes to assess and mitigate risk, and nurture a culture dedicated to curating AI responsibly. At the organizational level, AI governance decides and drives the AI strategy for an organization. Based on the AI strategy, the organization establishes AI principles and policies, and set up an AI Ethics Board which supports business rules and guidelines creation, and transparent reporting mechanisms. As part

of the AI Organizational Governance, appropriate guardrails and parameters are determined. These efforts mitigate risks and build processes to provide initial assessments and assist the organization in future self-assessments on its AI governance maturity level.

On the AI model operational level, governance is focused on the specific implementation, monitoring, and management of individual AI models through their lifecycle. This includes ongoing assessments of model performance, bias detection and mitigation, interpretability, transparency, and adherence to ethical guidelines during real-time operation. Operational governance is crucial for addressing emerging challenges and ensuring that AI systems align with the organization's values throughout their lifecycle. IBM watsonx.governance platform and IBM open-source 360 toolkits can automate the AI model governance, bringing efficiency and effectiveness to the practice.

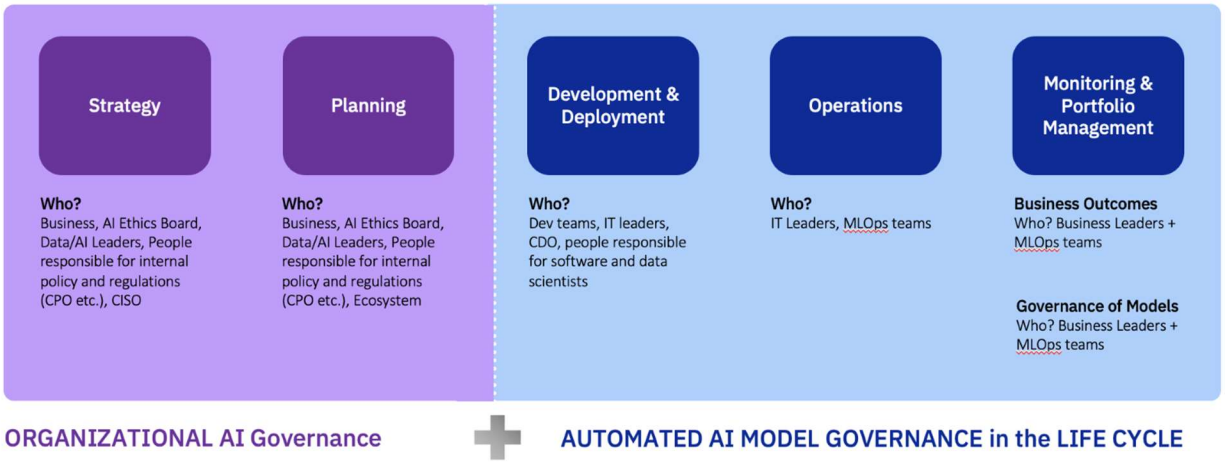

Figure 1. Comprehensive AI Governance Framework

ESTABLISHING ORGANIZATIONAL AI GOVERNANCE

When establishing AI governance at the organizational level, multi-disciplinary teams are essential for a holistic approach to this socio-technical challenge. The teams can use design thinking and garage methodologies, establish communication strategies, establish AI Ethics board, establish Trustworthy AI Center of Excellences, and drive education and diversity for teams. The multi-disciplinary teams shall establish processes to incorporate stakeholder perspectives, consider impacts to the enterprise and broader ecosystems, set AI and data risk profiles, establish an organizational structure, and assess and audit models. The multi-disciplinary teams shall identify requirements for privacy, robustness, fairness, explainability, transparency to capture, report and review data. They define integrated methodologies and toolkits, customize scaled data science methods, and help assess risk and mitigate it holistically.

There may be some challenges that include concerns over safeguard rails, desire to train the trainer on how to assess for unintended patterns in the models and outputs, ensuring all regulatory and policies such as the AI EO are embraced in the governance process and unintended outcomes of the AI models such as hallucinations and drifting are prevented. The Figure 2 depicts both key areas and crucial steps that must be taken to establish a robust and effective organizational AI governance framework. It highlights the integral role of people, processes, and technology in cultivating a comprehensive approach to AI governance. An IBM Institute Business Value (IBV) study illustrated how this framework was leveraged to effectively operationalize AI in a way that was repeatable, sustainable, and trusted for a regional bank 2]. Trainings and awareness allow teams to use the framework for systemic empathy well before code is written to consider risks and design safeguard rails for the model.

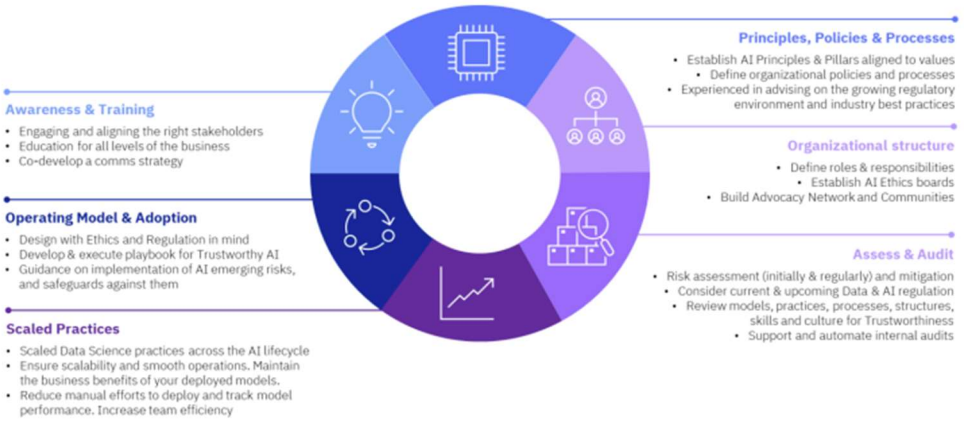

Figure 2. Key Areas and Crucial Steps for Establishing Organizational AI Governance

ASSESSING AI MODELS

The evaluation of AI models has traditionally focused on gauging their predictive power. However, instances of unintended consequences have prompted a recognition that there are other crucial dimensions to consider when evaluating AI models. For instance, a study published in Science in October 2019 has found that an algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care heavily favored white patients over black patients. While race itself wasn't a variable used in this algorithm, another variable highly correlated to race was, which was healthcare cost history. For various reasons, black patients incurred lower healthcare costs than white patients with the same conditions on average [3]. According to another study published in Science Advances in August 2022, diagnostic and decision support tools used in dermatology are AI models trained on a dataset of skin tones and skin conditions. However, the datasets the state-of-the-art dermatology AI models are trained on are limited due to the lack of images of uncommon diseases, as well as diverse skin tones, resulting in their substantial limitations on dark skin tones and uncommon diseases [4].

We advocate for a holistic approach to assessing AI models, encompassing the following five key dimensions: (i) Competence, (ii) Fairness, (iii) Robustness, (iv) Openness, and (v) Privacy. We will describe each of these five aspects and the open-source toolkits that IBM has developed to address them in the following section. IBM watsonx.governance is a platform that has incorporated these elements and employs software automation to strengthen an organization's ability to mitigate risks, manage regulatory requirements and address ethical concerns for both generative AI and machine learning (ML) models.

COMPETENCE

Competence refers to the basic performance of an AI/ML model, I.e., the accuracy of the model. Good performance, appropriately quantified based on the specifics of the problem and application, is a necessity to be used in any real-world task. Testing and validation of ML systems is different from testing traditional software systems [5]. Since the whole point of machine learning systems is to generalize from training data to label new unseen input data points, they suffer from the oracle problem: not knowing what the correct answer is supposed to be for a given input. Testing for accuracy (and related performance metrics like F1 score and AUC) is not sufficient when evaluating AI models as quantifying aleatoric (aka statistical) and (aka systematic) uncertainty around the test results is as important. You also need to test for accuracy under distribution shifts. There are inherent uncertainties in the predictions of ML models. Knowing how uncertain the prediction might be influences how people act on it.

Uncertainty Quantification 360 (https://uq360.res.ibm.com/) [6] is an open-source tool that IBM has developed to streamline the process of estimating, evaluating, improving, and communicating uncertainty of machine learning model performance through its lifecycle. The data scientists can use the tool to evaluate the uncertainties of the prediction, use the uncertainties to decide how to improve the model, and assist the decision making based on the uncertainties.

FAIRNESS

Although machine learning, by its very nature, is always a form of statistical discrimination, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage. Biases in training data, due to either prejudice in labels or under-/over-sampling, yields models with unwanted bias. Fairness can be measured at different points in a machine learning pipeline: on the training data or on the learned model, which also relates to the pre-processing, in-processing, and post-processing categories of bias mitigation algorithms.

AI Fairness 360 (https://aif360.res.ibm.com/) [7] is an open-source toolkit that IBM has developed to examine and evaluate discrimination and bias in AI models, including age bias, racial bias, etc., and provide mitigation recommendations. It includes a comprehensive set of metrics for datasets and models to test for bias and explanations for these metrics. It also includes bias mitigation algorithms to improve fairness metrics by modifying the training data, the learning algorithm, or the predictions even though we recommend the earliest mediation category in the pipeline that the user has permission to apply because it gives the most flexibility and opportunity to correct bias as much as possible. If possible, all algorithms from all permissible categories should be tested because the ultimate performance depends on dataset characteristics. Figure 3 shows an example where applying Reweighing Algorithm to the training data from the Compas recidivism prediction dataset in AI Fairness 360 reduces the level of bias between female and male measured by various bias metrics.

Figure 3. Comparing original vs. mitigated results by applying Reweighing Algorithm

ROBUSTNESS

Robustness includes safety and security of AI/ML models and systems. AI/ML systems need to maintain good and correct performance across varying operating conditions. Different conditions could come from natural changes in the world or from malevolent or benevolent human-induced changes.

Another open-source toolkit that IBM has developed, the Adversarial Robustness 360 (https://art360.res.ibm.com/) [8], can defend and verify AI/ML models against adversarial attacks. ART addresses growing concerns about people's trust in AI, specifically the security of AI in mission-critical applications. ART provides a comprehensive and growing set of tools to systematically assess and improve the robustness of AI models against adversarial attacks, including evasion, poisoning, extraction, and inference. For instance, the adversary might seek to add adversarial noise to an input and create an adversarial sample. These samples, when provided to a well-trained target model, cause predictable errors at the model's output. The adversary could also use direct or indirect methods to corrupt the training data in order to achieve a specific goal. In addition, the adversary could use API access to a target blackbox model in order to extract information about the training data or to learn the target model's parameters. Defending Machine Learning models involves certifying and verifying model robustness and model hardening with approaches such as pre-processing inputs, augmenting training data with adversarial samples, and leveraging runtime detection methods to flag any inputs that might have been modified by an adversary. Adversarial Robustness 360 can also be used to create adversarial attacks against Machine Learning models which is required to test defenses with state-of-the-art threat models.

OPENNESS

Openness applies to various aspects related to human interaction with AI/ML models. This includes communication from the machine to the human through interpretability and explainability of models, the system reporting its quantified uncertainty or confidence, and transparency into overall AI/ML system pipelines and lifecycles.

The open-source AI Explainability 360 (https://aix360.res.ibm.com/) [9], developed by IBM can help make AI/ML models understandable. It includes many different algorithms to explain: data vs. model, directly interpretable vs. post hoc explanation, local vs. global, static vs. interactive; the appropriate choice depends on the persona of the consumer of the explanation [10]. For the FICO Explainable Machine Learning Challenge, the fundamental task is to use the information about the applicant in their credit report to predict whether they will make timely payments over a two-year period. The machine learning prediction is then used by loan officers to decide whether the homeowner qualifies for a line of credit. AI Explainability 360 does not only provide a global view into how the model predicts repayment

probability, it also provides explanations on why the application by a specific applicant was rejected by identifying top factors and their relative importance of factors contributing to the denial, as shown in Figure 4.

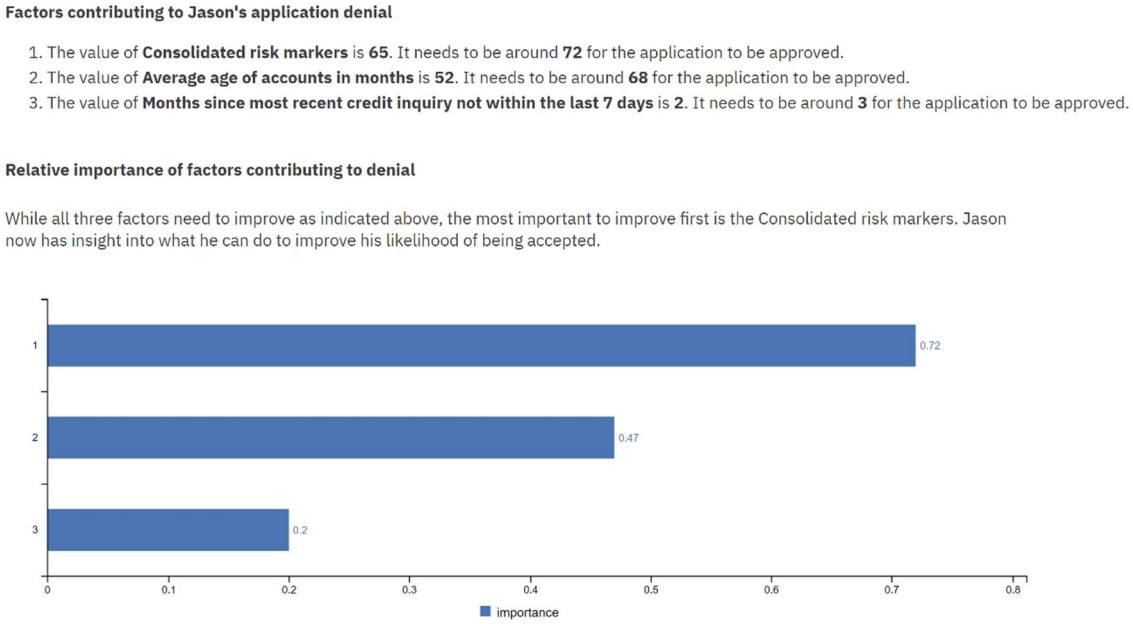

Figure 4. Factors contributing to a loan application denial and their relative importance

AI FactSheets are crucial to foster trust in AI by increasing transparency and enabling governance by providing information to AI consumers and regulators on how the AI model or service was created [11]. They are collections of relevant information (facts), which could range from information about the purpose and criticality of the model, measured characteristics of the dataset, model, or service, or actions taken during the creation and deployment process of the model or service. FactSheets are tailored to the particular AI model or service being documented and to the needs of their target audience or consumer, and thus can vary in content and format. FactSheets capture model or service facts from the entire AI lifecycle, and are compiled with inputs from multiple roles in this lifecycle as they perform their actions to increase the accuracy of these facts. In AI FactSheets 360 (https://aifs360.res.ibm.com/), IBM has developed a number of templates and a methodology for determining what template is appropriate for a particular model/service, use case, and target audience [12].

PRIVACY

AI systems must prioritize and safeguard consumers' privacy and data rights. Recent studies show that a malicious third party with access to a trained ML model, even without access to the training data itself, can still reveal sensitive, personal information about the people whose data was used to train the model. It is therefore crucial to be able to recognize and protect AI models that may contain personal information.

IBM's open-source tool, AI Privacy 360 (https://aip360.res.ibm.com/), can assess and protect against the privacy risks of the AI/M-enabled medical devices. The tool can also provide recommendations for adhering to any relevant privacy requirements by exploring tradeoffs between privacy, accuracy, and performance of the model. The tool continues to evolve to incorporate state-of-the-art techniques. In an example Random Forest model predicting mortality of hepatitis patients, the model accuracy was 0.8 and the attack accuracy was 0.56, determined by AI Privacy 360, which means that some membership information was leaked (the attack accuracy exceeds randomly guessed predictions of 0.5). After model anonymization is applied in AI Privacy 360, the attack accuracy dropped to 0.48 which means the privacy leakage was addressed while the model accuracy stays at a similar level, as shown in Table 1.

Table 1. Original model results vs. model results after privacy leakage stopped

	Model Accuracy	Attack Accuracy
Original	0.8	0.56
Mitigated	0.79	0.48

CONCLUSION

Every person involved in the creation of AI at any step is accountable for considering the system's impact in the world. Curating AI, developing, and deploying trustworthy in AI is not a technical problem with a technical solution. It is a socio-technical challenge that, to solve, requires a holistic approach encompassing people, processes and tools. There are elements to earn trust and ensue that AI models are accurate, auditable, explainable, fair, and protective of data privacy.

REFERENCES

1. IBM AI Governance Consulting. https://www.ibm.com/consulting/ai-governance
2. IBM Institute Business Value AI ethics in action Report. https://www.ibm.com/downloads/cas/4DPJK92W
3. Ziad Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations. Science 366,447-453(2019).
4. Roxana Daneshjou et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci. Adv.8, eabq6147(2022).
5. P. Santhanam, "Quality Management of Machine Learning Systems". In: Shehory, O., Farchi, E., Barash, G. (eds) Engineering Dependable and Secure Machine Learning Systems. EDSMLS (2020). Communications in Computer and Information Science, vol 1272. Springer, Cham. https://doi.org/10.1007/978-3-030-62144-5_1
6. S. Ghosh, Q. V. Liao, K. N. Ramamurthy, J. Navratil, P. Setiger, K. R. Varshney, and Y. Zhang. "Uncertainty Quantification 360." In Joint International Conference on Data Science & Management of Data, pp. 333-335. 2022.
7. R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." IBM Journal of Research and Development 63, no. 4/5 (2019): 4.
8. M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, et al. "Adversarial Robustness Toolbox v1. 0.0." arXiv:1807.01069 (2018).
9. V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, et al. "AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models." J. Mach. Learn. Res. 21, no. 130 (2020).
10. S. Dey, P. Chakraborty, B. C. Kwon, A. Dhurandhar, et al. "Human-centered explainability for life sciences, healthcare, and medical informatics." Patterns 3, no. 5 (2022): 100493.
11. M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, et al. "FactSheets: Increasing trust in AI services through supplier's declarations of conformity." IBM Journal of Research and Development 63, no. 4/5 (2019): 6.
12. J. Richards, D. Piorkowski, M. Hind, S. Houde, A. Mojsilovic, and K. R. Varshney. "A Human-Centered Methodology for Creating AI FactSheets." IEEE Data Eng. Bull. 44(4) (2021).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Qingying (Ally) Lu
Company: IBM
Address: 2300 Dulles Station Blvd., Suite 200, Herndon, VA 20171
Email: qingying.lu@us.ibm.com
Website: www.ibm.com