A Brief Overview of AI Governance for Responsible Machine Learning Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

Organizations of all sizes, across all industries and domains are leveraging artificial intelligence (AI) technologies to solve some of their biggest challenges around operations, customer experience, and much more. However, due to the probabilistic nature of AI, the risks associated with it are far greater than traditional technologies. Research has shown that these risks can range anywhere from regulatory, compliance, reputational, user trust and societal risks, to financial and even existential risks. Depending on the nature and size of the organization, AI technologies can pose a significant risk, if not used in a responsible way. This position paper seeks to present a brief introduction to AI governance, which is a framework designed to oversee the responsible use of AI with the goal of preventing and mitigating risks. Having such a framework will not only manage risks but also gain maximum value out of AI projects and develop consistency for organization-wide adoption of AI.

1 Introduction

In this position paper, we share our insights about AI Governance in companies, which enables new connections between various aspects and properties of trustworthy and socially responsible ML: security, robustness, privacy, fairness, ethics, interpretability, transparency, etc.

For a long time artificial intelligence (AI) was something enterprise organizations adopted due to the huge amounts of resources they have at their fingertips. Today, smaller companies are able to take advantage of AI due to newer technologies, e.g. cloud software, which are significantly more affordable than what was available in the past [2, 8, 11, 12]. AI has been on an upward trajectory in recent years and it will increase significantly over the next several years [21] [15] [28]. Obviously, AI is seen as a fruitful investment for many organizations. However, every investment has its pros and cons. Unfortunately, the cons associated with AI adoption are caused by the builders of such AI systems who do not take the necessary steps to avoid problems down the road [20, 12].

1.1 Problems within Industry

Applications of AI in industry is still in its infancy. With that said, many problems have arisen since its adoption [10, 12, 20], which can be attributed to several factors:

Lack of risk management: Too much attention is given to applications of AI and its potential success and not enough attention is given to its potential pitfalls & risks.

AI adoption is moving too fast: According to a survey by KPMG in 2021 [16], many respondents noted that AI technology is moving too fast for their comfort in industrial manufacturing (55%), technology (49%), financial services (37%), government (37%), and health care (35%) sectors.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

AI adoption needs government intervention: According to the same survey by KPMG, [16] an overwhelming percentage of respondents agreed that governments should be involved in regulating AI technology in the industrial manufacturing (94%), retail (87%), financial services (86%), life sciences (86%), technology (86%), health care (84%), and government (82%) sectors.

Companies are still immature when it comes to adopting AI: Some companies are not prepared for business conditions to change once a ML model is deployed into the real world.

Many of these problems can be avoided with proper governance mechanisms. AI without such mechanisms is a dangerous game with detrimental outcomes due its inherent uncertainty [36, 7]. With that said, adding governance into applications of AI is imperative to ensure safety in production.

2 AI Governance

In order for organizations to realize the maximum value out of AI projects and develop consistency for organization-wide adoption of AI, while managing significant risks to their business, they must implement AI governance [30, 9, 13, 17]; this enables organizations to not only develop AI projects in a responsible way, but also ensure that there is consistency across the entire organization and the business objective is front and center. With the AI governance implemented (as illustrated in Figure 1), the following benefits can be realized:

Alignment and Clarity: teams would be aware and aligned on what the industry, international, regional, local, and organizational policies are that need to be adhered to.

Thoughtfulness and Accountability: teams would put deliberate effort into justifying the business case for AI projects, and put conscious effort into thinking about end-user experience, adversarial impacts, public safety & privacy. This also places greater accountability on the teams developing their respective AI projects.

Consistency and Organizational Adoption: teams would have a more consistent way of developing and collaborating on their AI projects, leading to increased tracking and transparency for their projects. This also provides an overarching view of all AI projects going on within the organization, leading to increased visibility and overall adoption.

Process, Communication, and Tools: teams would have complete understanding of what the steps are in order to move the AI project to production to start realizing business value. They would also be able to leverage tools that take them through the defined process, while being able to communicate with the right stakeholders through the tool.

Trust and Public Perception: as teams build out their AI projects more thoughtfully, this will inherently build trust amongst customers and end users, and therefore a positive public perception.

The components of AI governance should focus on organizational and use case planning, AI development, and AI "operationalization", which come together to make a **4 stage AI life-cycle approach**.

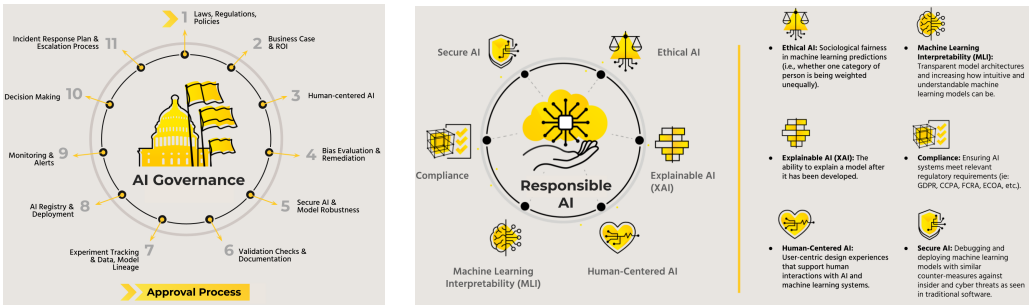

Figure 1: Illustration of the AI Governance application towards responsible AI in companies.

2.1 Organizational Planning

An AI Governance Program [17, 9, 35] should be organized in such a way that (a) there is comprehensive understanding of regulations, laws, and policies amongst all team members (b) resources and help available for team members who encounter challenges (c) there is a light weight, yet clear process to assist team members.

1. **Regulations, Laws, Policies**

   Laws and regulations that apply to a specific entity should be identified, documented, and available for others to review and audit. These regulations, laws, and policies vary across industry and sometimes by geographical location. Organizations should, if applicable, develop policies for themselves, which reflect their values and ethical views [35, 24, 12]; this enables teams to be more autonomous and make decisions with confidence.

2. **Organization (Center of Competency)**

   Establishing groups within an organization that provide support to teams with AI projects can prove to be quite beneficial. This includes a group that is knowledgeable with regulations, laws, and policies and can answer any questions that AI teams may have; a group that is able to share best practices across different AI teams within the organization; a group that is able to audit the data, model, process, etc. to ensure there isn't a breach or non-compliance. For more information, we refer the reader to to the survey by Floridi *et al.* [12].

3. **Process**

   Developing a light-weight process that provides guidelines to AI teams can help with their efficiency, rather than hinder their progress and velocity. This involves identifying what the approval process and incident response would be for data, model, deployments, etc.

2.2 Use Case Planning

Building use cases involves establishing business value, technology stack, and model usage. The group of people involved in this process can include: subject matter experts, data scientists/analysts/annotators and ML engineers, IT professionals, and finance departments.

Business Value Framework. The AI team should ensure that the motivation for the AI use case is documented and communicated amongst all stakeholders. This should also include the original hypothesis, and the metrics that would be used for evaluating the experiments.

Tools, Technology, Products. The AI team should either select from a set of pre-approved tools and products from the organization or get a set of tools and products approved before using in an AI user case. If tools for AI development are not governed, it not only leads to high costs and inability to manage the tools (as existing IT teams are aware), it also leads to not being able to create repeatability and traceability into AI models.

Model Usage. Once a sense of value is attached to the use case, then the next step would be to break down the use case to its sub-components which include, but are not limited to, identifying the consumer of the model, the model's limitations, and potential bias that may exist within the model, along with its implications. Also, one would want to ensure inclusiveness of the target, public safety/user privacy, and identification of the model interface needed for their intended use case.

2.3 AI Development

Development of a machine learning model, including data handling and analysis, modeling, generating explanations, bias detection, accuracy and efficacy analysis, security and robustness checks, model lineage, validation, and documentation.

1. **Data Handling, Analysis and Modeling**

   The first technical step to any AI project is the procurement and analysis of data, which is critical as it lays the foundation for all work going forward. Once data is analyzed, then one must decipher if modeling is needed for the use case at hand. If modeling is needed, then the application of AI can take place. Such an application is an iterative process spanned across many different types of people.

2. **Explanations and Bias**

   The goal of model explanations is to relate feature values to model predictions in a human-friendly manner [25]. What one does with these explanations breaks down to 3 personas: modeler, intermediary user, and the end user. The modeler would use explanations for model debugging and gaining understanding of the model they just built. The intermediary user would use what the modeler made for actionable insights. And finally, the end user is the person the model affects directly. For these reasons, Explainable Artificial Intelligence (**XAI**) is a very active research topic [4, 14].

   Bias, whether intentional (disparate treatment) or unintentional (disparate impact), is a cause for concern in many applications of AI [31, 22]. Common things to investigate when it comes to preventing bias include the data source used for the modeling process, performance issues amongst different demographics, disparate impact, identifying known limitations & potential adverse implications, and the models impact on public safety [5]. We refer the reader to the survey by Mehrabi *et al.* [22] for more details about bias and fairness in AI.

3. **Accuracy, Efficacy, & Robustness**

   Accuracy of a machine learning model is critical for any business application in which predictions drive potential actions. However, it is not the most important metric to optimize. One must also consider the efficacy of a model, i.e., is the model making the intended business impact?

   When a model is serving predictions in a production setting, the data can be a little or significantly different from the data that the project team had access to. Although model drift and feature drift can capture this discrepancy, it is a lagging indicator, and by that time, the model has already made predictions. This is where **Robustness** comes in: project teams can proactively test for model robustness, using "out of scope" data, to understand whether the model perturbs. The out of scope data can be a combination of manual generation (toggle with feature values) and automatic generation (system toggles feature values).

4. **Security**

   ML systems today are subject to general attacks that can affect any public facing IT system [6], [27]; specialized attacks that exploit insider access to data and ML code; external access to ML prediction APIs and endpoints [33], [32]; and trojans that can hide in third-party ML artifacts. Such attacks must be accounted for and tested against before sending a machine learning model out into the real world.

5. **Documentation & Validation**

   An overall lineage of the entire AI project life-cycle should be documented to ensure transparency and understanding [23], [29], which will be useful for the AI team working on the project and also future teams who must reference this project for their own application.

   Model validation [26, 18] is the set of processes and activities that are carried out by a third party, with the intent to verify that models are robust and performing as expected, in line with the business use case. It also identifies the impact of potential limitations and assumptions. From a technical standpoint, the following should be considered: (i) Sensitivity Analysis. (ii) In-sample vs. Out-of-sample performance. (iii) Replication of results from model development team. (iv) Stability analysis. Model "validators" should document all of their findings and share with relevant stakeholders.

2.4 AI Operationalization

Deploying a machine learning model into production (*i.e.* MLOps [1, 34]) is the first step to potentially receiving value out of it. The steps that go into the deployment process should include the following:

Review-Approval Flow: Model building in an AI project will go through various stages: experimentation, model registration, deployment, and decommissioning. Moving from one stage to the next would require "external" reviewer(s) who will vet and provide feedback.

Monitoring & Alerts: Once a model is deployed, it must be monitored for various metrics to ensure there is not any degradation in the model. The cause for a model degrading when deployed can include

the following: feature and/or target drift, lack of data integrity, and outliers, amongst other things. In terms of monitoring, accuracy, fairness, and explanations of predictions are of interest [3, 19].

Decision Making: The output of a machine learning model is a prediction, but that output must be turned into a decision. How to decide? Will it be autonomous? Will it involve a human in the loop? The answers to these questions vary across different applications, but the idea remains the same, ensuring decisions are made in the proper way to decrease risk for everyone involved.

Incident Response and Escalation Process: With AI models being used in production, there is always going to be a chance for issues to arise. Organizations should have an incident response plan and escalation process documented and known to all project teams.

3 Conclusion

AI systems are used today to make life-altering decisions about employment, bail, parole, and lending, and the scope of decisions delegated by AI systems seems likely to expand in the future. The pervasiveness of AI across many fields is something that will not slowdown anytime soon and organizations will want to keep up with such applications. However, they must be cognisant of the risks that come with AI and have guidelines around how they approach applications of AI to avoid such risks. By establishing a framework for AI Governance, organizations will be able to harness AI for their use cases while at the same time avoiding risks and having plans in place for risk mitigation, which is paramount.

Social Impact As we discuss in this paper, governance and certain control over AI applications in organizations should be mandatory. This "new term", *AI Governance*, aims to enable and facilitate connections between various aspects of trustworthy and socially responsible Machine Learning systems, and therefore it accounts for security, robustness, privacy, fairness, ethics, and transparency. We here summarize this concept, and we believe the implementation of these ideas should have a positive impact in the society.

References

[1] Sridhar Alla and Suman Kalyan Adari. What is mlops? In *Beginning MLOps with MLFlow*, pages 79–124. Springer, 2021.

[2] Sulaiman AlSheibani, Yen Cheung, and Chris Messom. Artificial intelligence adoption: Ai-readiness at firm-level. *PACIS 2018 Proceedings*, 2018.

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

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[5] S. Barocas, M. Hardt, and A. Narayanan. Fairness and Machine Learning: Limitaions and Opportunities, 2022. URL: `https://fairmlbook.org/pdf/fairmlbook.pdf`.

[6] Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The Security of Machine Learning. *Machine Learning*, 81(2):121–148, 2010. URL: `https://people.eecs.berkeley.edu/~adj/publications/paper-files/SecML-MLJ2010.pdf`.

[7] John Bresina, Richard Dearden, Nicolas Meuleau, Sailesh Ramkrishnan, David Smith, and Richard Washington. Planning under continuous time and resource uncertainty: A challenge for ai. *arXiv preprint arXiv:1301.0559*, 2012.

[8] Marija Cubric. Drivers, barriers and social considerations for ai adoption in business and management: A tertiary study. *Technology in Society*, 62:101257, 2020.

[9] Allan Dafoe. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

[10] Artificial Intelligence Incident Database. AI Incident Database, 2022. URL: https://incidentdatabase.ai/.

[11] Yanqing Duan, John S Edwards, and Yogesh K Dwivedi. Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. *International journal of information management*, 48:63–71, 2019.

[12] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4): 689–707, 2018.

[13] Urs Gasser and Virgilio AF Almeida. A layered model for ai governance. *IEEE Internet Computing*, 21(6): 58–62, 2017.

[14] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019.

[15] IBM. IBM Global AI Adoption Index, 2022. URL: https://www.ibm.com/downloads/cas/GVAGA3JP.

[16] KPMG. AI adoption accelerated during the pandemic but many say it's moving too fast: KPMG survey, 2021. URL: https://info.kpmg.us/news-perspectives/technology-innovation/thriving-in-an-ai-world/ai-adoption-accelerated-during-pandemic.html.

[17] Maciej Kuziemski and Gianluca Misuraca. Ai governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy*, 44(6):101976, 2020.

[18] Maurice Landry, Jean-Louis Malouin, and Muhittin Oral. Model validation in operations research. *European journal of operational research*, 14(3):207–220, 1983.

[19] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[20] Bernard Marr. Is artificial intelligence dangerous? 6 ai risks everyone should know about. *Forbes*, 2018.

[21] Joe McKendrick. AI Adoption Skyrocketed Over the Last 18 Months. *Harvard Business Review*, 2021. URL: https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months.

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[23] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145%2F3287560.3287596.

[24] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11): 501–507, 2019.

[25] Christopher Molnar. *Interpretable Machine Learning*. 2022. URL: https://christophm.github.io/interpretable-ml-book/.

[26] Board of Governors of the Federal Reserve System Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management, SR Letter 11-7, 2011. URL https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

[27] Nicolas Papernot. A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. ACM, 2018. URL: https://arxiv.org/pdf/1811.01134.pdf.

[28] PwC (PriceWaterhouseCoopers). PwC 2022 AI Business Survey, 2022. URL: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html.

[29] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai, 2022. URL https://arxiv.org/abs/2204.01075.

[30] Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27(3):491–497, 2020.

[31] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a standard for identifying and managing bias in artificial intelligence, 2022-03-15 04:03:00 2022. URL `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464`.

[32] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. URL: `https://arxiv.org/pdf/1610.05820.pdf`.

[33] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016. URL: `https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf`.

[34] Mark Treveil, Nicolas Omont, Clément Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoillotte, Makoto Miyazaki, and Lynn Heidmann. *Introducing MLOps*. O'Reilly Media, 2020.

[35] Weiyu Wang and Keng Siau. Artificial intelligence: A study on governance, policies, and regulations. *MWAIS 2018 proceedings*, 40, 2018.

[36] Lotfi A Zadeh. Is probability theory sufficient for dealing with uncertainty in ai: A negative view. In *Machine intelligence and pattern recognition*, volume 4, pages 103–116. Elsevier, 1986.