Managing AI: Risks and Opportunities

Donald R. Polaski[1] and Marissa J. Brienza[2]

[1]Booz Allen Hamilton, 8283 Greensboro Drive, Hamilton Building, McLean, VA 22102 USA; polaski_donald@bah.com

[2]Booz Allen Hamilton, 8283 Greensboro Drive, Hamilton Building, McLean, VA 22102 USA; brienza_marissa@bah.com

ABSTRACT

Artificial intelligence (AI) is fundamentally changing the way humans interact with machines by automating tasks which before only humans could perform. While AI can seem like magic, using these innovative techniques comes with considerable risks. AI models can be fraught with bias, as was the case when Amazon launched an internal recruiting tool that used AI to vet job resumes. In designing the tool, researchers identified that the model was ranking women's resumes significantly lower than men's resumes. The model penalized resumes for having the term "women" in activities like "women's chess club captain" and downgraded applicants for having attended all-women's colleges. While Amazon scrapped the program in 2015, it is an important lesson that even organizations with the best intentions may run into unexpected risks when managing AI-based projects. In this paper we review multiple case studies of how AI projects realized risk and highlight additional risks associated with managing AI projects. Topics addressed include bias in AI models, privacy concerns with data used to train AI systems, legal issues that may rise from using AI-powered tools, lack of model transparency and explainability, and model drift – the concept of AI models losing accuracy over time. We also cover strategies for dealing with these risks to help program managers maximize the impact that AI has on their projects.

INTRODUCTION

Artificial intelligence (AI) is fundamentally changing the way humans interact with machines by automating tasks which before only humans could perform. While AI can seem like magic, using these innovative techniques comes with considerable risks. AI models can be fraught with bias, as was the case when Amazon launched an internal recruiting tool that used AI to vet job resumes. In designing the tool, researchers identified that the model was ranking women's resumes significantly lower than men's resumes. The model penalized resumes for having the term "women" in activities like "women's chess club captain" and downgraded applicants for having attended all-women's colleges. While Amazon scrapped the program in 2015, it is an important lesson that even organizations with the best intentions may run into unexpected risks when managing AI-based projects. In this paper we review multiple case studies of how AI projects realized risk and highlight additional risks associated with managing AI projects. Topics addressed include bias in AI models, privacy concerns with data used to train AI systems, legal issues that may rise

from using AI-powered tools, lack of model transparency and explainability, and model drift – the concept of AI models losing accuracy over time. We also cover strategies for dealing with these risks to help program managers maximize the impact that AI has on their projects.

AI BIAS

AI bias occurs when an AI system produces outputs that lead to discrimination against specific groups or individuals (Belenguer, 2022). PMs should be concerned about bias in the AI tools they use or develop because biased AI can lead to unfair and discriminatory outcomes. If the AI tools are used in critical areas such as hiring, lending, or criminal justice, biased outcomes can have dire consequences for individuals and society.

Take for example an AI project to assist recruiters and managers in hiring decisions developed and ultimately shelved by Amazon. As early as 2014, Amazon had been building AI systems to review job applicants resumes and assign a one to five star ranking to help separate strong candidates from weak candidates (Goodman, 2018). On its surface, using AI to filter out candidates could result in real time savings for an enterprise as large as Amazon. In execution, the company quickly identified that their AI system had built in bias against female candidates. Resumes that included the term “women” as in “women’s chess club captain” would be downgrade. The AI system also downgraded graduates of two women’s colleges. (Dustin, 2018). Part of the reason this bias was present in the recruiting tool was due to an imbalance in the data used to train it. The tech industry is overwhelmingly male (Hupfer, 2021) and as such the resumes used to train the underlying algorithms came primarily from male candidates. Because Amazon hired mostly men in the past, men scored higher in their new recruiting tool. Fortunately, Amazon was quick to catch on to the bias in their system and terminated the program. Unfortunately, this kind of bias can creep into any AI system where historical bias may be present in the data used to train it.

To reduce the risk of AI bias program managers can ask the following questions during project execution:

- Where did the training data come from for the AI models we are building/AI tools we are using? Is there reason to believe biases or inaccuracies exist in the training data?
- Did our team ensure proper demographic representation in data used to train our AI models/tools?
- Have the models/tools we are using been tested across a diverse set of data to ensure it performs appropriately across all demographics/groups?
- What tools do we have in place to monitor and evaluate our model’s continued performance to ensure bias does not enter the system over time?

DATA PRIVACY CONCERNS

Training AI models requires large volumes of data. As organizations accumulate more data to stand up AI programs, safeguarding the data, ensuring data storage meets regulatory standards, and protecting the privacy of the users generating that data becomes increasingly important. Failure to follow applicable data privacy law could result in significant fines, lawsuits, or even the prohibition of a product in certain countries. Data privacy concerns can emerge in multiple ways on any given project:

- **Data Security:** Storing large amounts of data creates an inherent risk that nefarious actors may retrieve the stored information via a data breach. If sensitive data like personal identifiable information (PII) or protected health information (PHI) is compromised, it can result in reputational harm to the organization and significant harm to individuals. (IBM Security, 2022)
- **Data Repurposing:** Organizations may collect data for one purpose such as marketing and then repurpose that data to train AI models. This can result in privacy violations as data may end up used in ways that users did not consent to at the time of collection. (Privacy Commissioner of Canada, 2021)
- **Data Re-Identification:** Even data that has PII and PHI removed can be used to identify specific individuals with high levels of accuracy (Rocher, 2019).

An incident involving Google and the National Institute of Health (NIH) provides a sobering example of how data privacy concerns can derail an AI project. In 2017 Google pulled out of a project with the NIH just days before more than 100,000 chest X-rays were to be publicly posted to the internet. The decision was made based on concerns that the images could be used to identify patients (MacMillan, 2019) Googled planned to host the images on their cloud servers and make them available to demonstrate how machine learning tools like Google's TensorFlow library could be used to identify lung disease. While the images were anonymized by Google and NIH staff, the NIH later determined that dozens of images still included PII including the dates of the x-ray and images of distinctive jewelry that patients wore during the x-rays. Due to legal and privacy concerns, Google opted to terminate the program to minimize their risk exposure.

To help avoid data security risks, PMs should document the answers to the following questions

- What regulations for safeguarding data exist within my industry, state, or country?
- How is my team ensuring that we meet these regulations throughout the AI development lifecycle?
- What data governance exists in my company, and what requirements do we need to meet throughout the data management lifecycle?
- How might nefarious actors use our data for re-identification and how do we anonymize the data fully to prevent this outcome?
- Is our data encrypted, and have we restricted access to the data to ensure that only those with the need to access it can access it?
- How are we obtaining consent from individuals to use their data?
- What mechanisms are in place to prevent, detect, and address security breaches?

GENERATIVE AI – LEGAL RISKS

One of the most exciting developments in AI over the last few years is generative AI. Generative AI can create new data including images, videos, music, and text based on patterns and structures found in existing data. The most famous example of generative AI is ChatGPT. Developed by OpenAI, ChatGPT allows users to write jokes, essays, songs, poetry, and software by giving it commands like "write a poem about managing AI risk in the style of T.S. Elliott" or "tell me a joke about program management that includes a giraffe" (McKinsey, 2023). The full power of ChatGPT is still being understood, but its popularity is unquestionable. From November 2022 to January 2023, ChatGPT increased its userbase to 100M active users per month (Hu, 2023). DALL-E provides another example of generative AI where users can provide written prompts to generate images in the style of their favorite artists. In both cases, AI researchers built these models by ingesting millions of pieces of data collected from across the internet.

ChatGPT In Action – A PM Joke

Input to ChatGPT: *"Tell me a joke about program management that includes a giraffe"*

ChatGPT's Response: "*Why did the giraffe become a program manager? Because they had a great overview of the project!"* (ChatGPT, personal communication, March 10, 2023)

While generative AI has the potential to fundamentally change the creative process there are multiple risks that PMs should be aware of before using the technology on their projects. As an example, consider GitHub Copilot, a software tool developed by Microsoft and OpenAI that uses generative AI to suggest new code and entire functions in software development tools. Initial benchmarks show Copilot can increase software writing speed by 55% percent, which could significantly accelerate project timelines (Kalliamvakou, 2022). GitHub created Copilot by ingesting billions of lines of computer code available on the internet (Ziegler, 2021). This approach has opened Copilot's creators to a class action lawsuit filed by Matthew Butterick and the Joseph Svaeri Law Firm (Vincent, 2022). The lawsuit contends that by training their AI system on public GitHub repositories, the defendants have violated the legal rights of a vast number of creators who posted their code under certain open-source licenses that require attribution and inclusion of the author's name and copyright in any derivative product (e.g., Apache license, Gnu Public License, MIT license) (Butterick, 2022). To further complicate matters, Copilot users have used the tool to generate blocks of copyrighted code, with no attribution and without attaching the required license.

While the class action lawsuit may take years to work its way through the courts, PMs using generative AI/ML in their projects should continue to monitor the cases progress. In a worst-case scenario, integrating AI-generated code into a product without proper attribution and without complying with the code's original license and copyright provisions could make the product illegal to distribute. In addition, using tools like Copilot could open the product or company up to legal action from the code's original copyright holders. While PMs must consider this risk when using generative AI tools, they must also work within their organization to understand the overall risk tolerance for incorporating generative AI. Until a legal precedent is established for these technologies, they will continue to operate in grey area that can create uncertainty for businesses and teams. This makes it even more important for PMs to prioritize transparency and ethical considerations when using generative AI.

MODEL TRANSPARENCY & EXPLAINABILITY

A common criticism of some AI systems is that they are "black boxes." Many popular AI models, such as deep learning models, learn patterns and insights by processing vast amounts of data and applying complex mathematical algorithms. This can result in highly performing AI systems that are difficult for humans to interpret. Unlike traditional systems that rely on explicit business rules and decision-making criteria, AI uses statistical methods not easily explained. This ambiguity has made it challenging to adopt AI in sensitive domains such as national defense and healthcare (Linardatos, 2021).

To better understand the risks associated with an AI model that lacks explainability, it is helpful to consider an example related to computer vision. One of the most popular examples concerns a model designed to distinguish between photos of wolves and huskies. In a 2016 experiment, Marco Riberio and his team trained an AI model by processing twenty labeled pictures where each wolf picture had snow in the background and each huskie picture did not (Riberio, 2016). The research team applied the model to a collection of sixty additional test images. Because of the initial bias in the training data, the model predicted “wolf” for every test image with snow and the model predicted “husky” for all other images, regardless of the animal’s color, markings, features, or position. Because every wolf picture also had snow in the training, the AI model had incorrectly determined that it was the presence of snow that differentiated wolves from huskies which lead to inaccuracy once it saw the additional test images.

Explainable AI (XAI) continues to be an area of highly active research as scientists and engineers strive to develop understandable AI systems trusted by their users (Saeed, 2023). While the specific techniques used to probe how a model makes decisions depends on the statistical methods used to build the model, working with your technical team to understand the answers to these questions will help to minimize the risk associated with building and fielding “black box” AI systems:

- What checks do we have in place to ensure that our models are not learning the wrong features due to bias in the training data set?
- What technologies and techniques are using to interrogate how our models are making predictions?
- Can we afford to reduce model accuracy to increase model transparency for our use case?
- How do we measure accuracy and performance of our AI system?

MODEL DRIFT

A common misconception when it comes to building solutions powered by AI is that once engineers train, validate, and deploy the underlying model it will continue to maintain the same level of performance over time. In reality, as soon as a team deploys an AI/ML model the accuracy and effectiveness of the model begins to deteriorate. Model drift can occur for a number of reasons. The most common cause is changes in the input data. For example, consider a model trained to scan emails and flag the ones that contain spam. Over time, the scammers sending those emails will adapt their tactics to avoid the AI algorithms classifying their emails as spam. If the spam detecting model never retrains on these new tactics, it will continue to perform worse than the day it was first deployed. Similarly, consider a model used to provide recommendations to a user (e.g., recommended products or TV shows to watch). If the model does not retrain as the user's tastes and preferences evolve it will continue to get worse over time reducing the likelihood of the user selecting its recommendations.

Model drift can have significant real-world consequences. Zillow, a website providing a real estate marketplace for buyers, sellers, and renters incorporated AI into one of their flagship products, Zillow Offers (Datta, 2021). Zillow Offers allows homeowners in certain markets to sell their homes quickly without the need for a traditional real estate agent. The product accomplished this by using AI to quickly assess a property's value based on multiple factors including home location, size, age, and condition. (Metz, 2021). Zillow followed recommended best practices for building its AI valuation model – rigorously testing models during development, operating the models for a pilot period, and rolling the model out slowly to see how it would perform in the real world. After reporting initial successes, the product team rapidly scaled the model to expand Zillow's purchasing program. In doing so, they purchased more homes in a six-month period than in the previous two years.

During this period the real-estate market, and thus the environment the model was meant to simulate, was rapidly changing. While we don't know exactly what went wrong with Zillow's approach, the consensus view is that Zillow overestimated the value of the homes they purchased because their algorithms did not adjust to market condition's rapidly enough (Editorial Team, 2021) The end result was $304M in losses due to the need to sell many of their purchased houses below their original purchase price.

While model drift can come from many sources, asking the following questions and understanding their answers will help to avoid realization of the risks outlined above:

- How are we monitoring model performance over time? What procedures are in place to alert engineers and project staff if models fall outside of expected performance bounds? What tools is the team using to accomplish this?
- How often do we retrain our models? Does this cadence make sense given the environment we are trying to make predictions in?
- How are we doing data quality control? If the data we feed into our models changes significantly, what controls do we have in place to identify those changes and alert the team?
- When integrating third party AI/ML technology, the technology vendor should address these questions. If the vendor does not have an adequate plan for dealing with model drift, it increases the probability of realizing model drift risks.

TOOLS FOR PROJECT MANAGERS

Industry and federal organizations have made progress in the last five years in creating tools to enable PMs to holistically address AI risk in their projects. In this section, we introduce some of those tools and resources and encourage PMs consider incorporating them into their project management processes.

Training: We recommend that all PMs working on AI-related projects develop a base literacy in the history, terminology, and high-level technical approaches of AI. While there are many online and university classes that can introduce AI, we recommend the on-demand training provided by Nvidia's Deep Learning Institute. Their Deep Learning Demystified course provides attendees with information regarding the history of AI and ML, discusses challenges organizations might face when adopting AI technology, introduces the latest tools and technologies associated with AI, and provides a roadmap of other training resources to continue an AI education. The course requires no prerequisites and is suitable for non-technical staff. At time of printing, this course could be found at the following link: https://www.nvidia.com/en-us/on-demand/session/gtcfall20-a21323eu/

National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF): In January 2023, NIST released the first version of their AI risk management framework, RMF 1.0 (NIST, 2023). Directed by Congress, NIST developed the AI RMF "to be used by organizations in varying degrees and capacities so that society can benefit from AI technologies while also being protected from its potential harms." The framework describes the characteristics of trustworthy AI systems, including ensuring the systems are valid and reliable, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias adequately managed. The framework enables organizations and individuals to jump start their approach to AI risk management by describing specific functions to help organizations address the risk of AI systems. Table 1 documents these functions.

Table 1. The Four Functions of the NIST AI RMF.

Function	Description
Govern	Focused on creating a culture of AI risk management within organizations; outlines processes, documents, and organizational schemes to manage AI risks; provides structure to align RMF with organizational value, principles, and policies; addresses legal and other issues concerning use of third-party software, hardware, and data within AI systems
Map	Focused on mapping the full end-to-end AI lifecycle to ensure that the interdependencies between elements of an AI project are understood; ensures AI risk management decisions at one stage of the AI lifecycle are not undermined by interactions and decisions at later stages of the AI activity; acts as input to the Measure and Manage functions of the RMF.
Measure	Employs quantitative and qualitative techniques and methodologies to analyze, benchmark, and monitor AI risk; Ensures AI systems are tested before deployment; Documents the metrics, processes, and procedures used to ensure validation is objective and repeatable.
Manage	Establishes plans for prioritizing risk and regular monitoring of AI systems; Employs systematic documentation approach established in the Govern function; Develops strategies to maximize AI benefit while minimizing negative impacts; Manages risks and benefits from third-party entities.

Understanding each function of the AI RMF, will help PMs understand how their work fits into their organization's overall AI strategy. However, for managing the risks of an AI project day-to-day, focusing on the 'Manage' function will provide the most value. In addition to the RMF itself, NIST has published an AI Risk Management Framework Playbook (RMF-P). Interested PMs can use the playbook to identify suggested actions, recommendations for documentation and transparency, and references to other documents with relevant information. As an example, the RMF-P provides a Responsible AI Impact Template Assessment to use in evaluating the risks associated with AI projects. By familiarizing themselves with the NIST AI RMF and the resources in the AI RMF-P, PMs can accelerate their understanding and adoption of the processes and tools required to manage risk in AI projects. At the time of this publication, the latest version of the NIST RMF and associated play book could be found at these URLs (https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, https://pages.nist.gov/AIRMF/)

REFERENCES

https://githubcopilotlitigation.com/

Belenguer, L. (2022). AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, 2, 771-787. https://link.springer.com/article/10.1007/s43681-022-00138-8#Abs1

Butterick, M. (2022, November 3). We've filed a lawsuit challenging GitHub Copilot, an AI product that relies on unprecedented open-source software piracy. Because AI needs to be fair & ethical for everyone. GitHub Copilot Litigation. https://githubcopilotlitigation.com/

Datta, A. (2021, October 20). The Dangers of AI Model Drift: Lessons to be Learned from the Case of Zillow Offers. https://aijourn.com/the-dangers-of-ai-model-drift-lessons-to-be-learned-from-the-case-of-zillow-offers/

Dustin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Goodman, R. (2018, October 10). Why Amazon's automated hiring tool discriminated against women. *ACLU*. https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against

Hu, K. (2018). ChatGPT sets record for fastest-growing user base – analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hupfer, S. (2021). Women in the tech industry: Gaining ground, but facing new headwinds. *Deloitte Insights.* https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/statistics-show-women-in-technology-are-facing-new-headwinds.html

IBM Security. (2022). *Cost of a Data Breach Report 2022* [White paper]. IBM. https://www.ibm.com/downloads/cas/3R8N1DZJ

Kalliamvakou, E. (2022, September 7). Research: quantifying GitHub Copilot's impact on developer productivity and happiness. https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-develop

er-productivity-and-happiness/
Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018
MacMillan, D., & Bensinger, G. (2019, November 15). Google almost made 100,000 chest X-rays public — until it realized personal data could be exposed. The Washington Post. https://www.washingtonpost.com/technology/2019/11/15/google-almost-made-chest-x-rays-public-until-it-realized-personal-data-could-be-exposed/
McKinsey & Company. (2021, October 19). What is generative AI? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai#/
Metz, C. (2022, November 23). Lawsuit Takes Aim at the Way A.I. Is Built. The New York Times. https://www.nytimes.com/2022/11/23/technology/copilot-microsoft-ai-lawsuit.html
Metz, R. (2021, November 9). Zillow's home-buying debacle shows how hard it is to use AI to value real estate. CNN Business. https://www.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html
National Institute of Standards and Technology. (2023, January 26). NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence [Press release]. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial
Privacy Commissioner of Canada. (2021). *Joint investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta.* OPC Actions and Decisions. https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/#toc2
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv preprint arXiv:1602.04938.
Rocher, L., Hendrickx, J.M., & de Montjoye, Y.A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), 3069. https://www.nature.com/articles/s41467-019-10933-3/
Ziegler, A. (2021, June 30). GitHub Copilot research recitation. GitHub Blog. https://github.blog/2021-06-30-github-copilot-research-recitation/