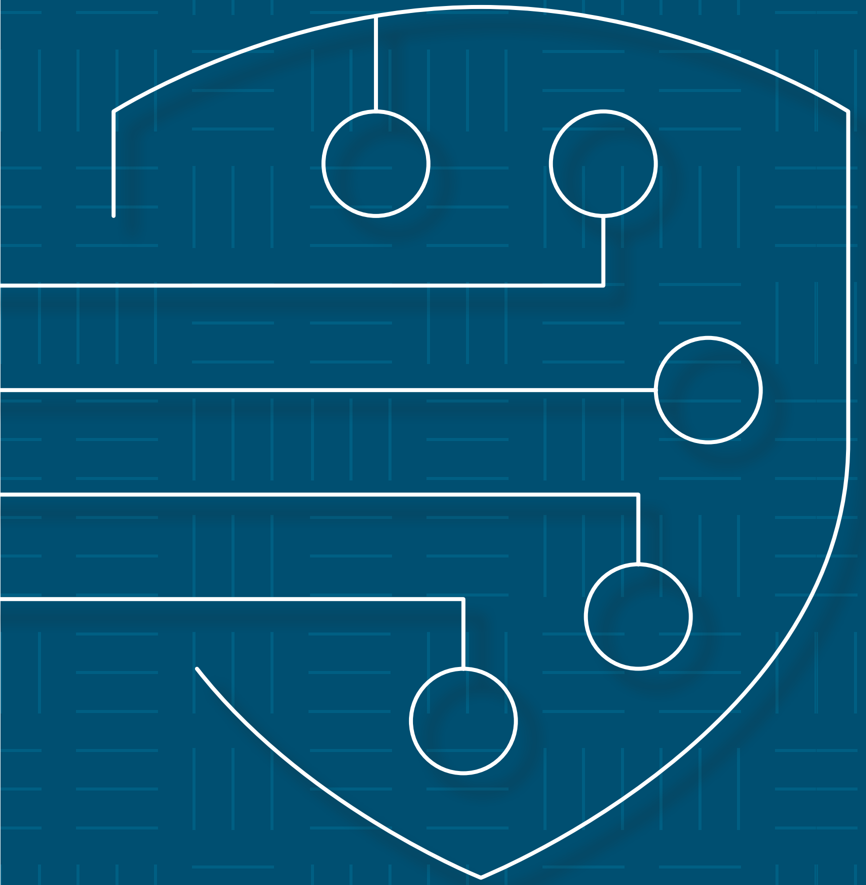
CompTIA
Artificial Intelligence Advisory Council

WHITE PAPER

Ethically Operationalizing AI and Machine Learning

Things to consider in order to minimize business consequences and liabilities



Created by CompTIA's

Artificial Intelligence
Advisory Council

Table of Contents:

Overview

People looking to deploy artificial intelligence or machine learning to a new business opportunity in their organization obviously look to characterize the benefits, costs, and risks. The ethical aspects of operationalizing AI may be more apparent in some domains that have been widely discussed and obviously egregious, such as a face recognition model that has poor performance on people with darker skin, a parole decision model that bases a large part of its recommendation on race that extends or amplifies existing biases, or a chatbot that occasionally spews sexist remarks. However, ethical considerations can emerge in almost any AI deployment in an organization.

Ethical considerations can add complexity to questions that may seem, at first glance, easy for a model to answer. Imagine creating a system designed to assess whether the driver of a vehicle should receive a ticket for exceeding the speed limit. Is a ticket issued every time the speed limit is exceeded? If the speed oscillates, is that multiple tickets? When does behavior become two speeding tickets? Should a ticket be issued if the driver speeds to avoid a collision or to drive someone to the emergency room?

Further, ethical considerations can appear in many more AI deployments than one might first imagine, even when the AI is making mundane decisions. If you're building a model to choose the location of a new franchise, what might the impact be on the local community in the long-term? If the model is based on historical decisions, is it amplifying biases in existing decisions from previous and current employees, that may become more obvious after use? If you're building a model to set the prices of products, treading the line between competitive pricing and maximizing revenue, is your model targeting higher prices toward people with lower incomes and less time because they have less time to research? Could the model make decisions that could drive away customers in the future because it treats customers unfairly? What if it accurately predicts something a customer thought was private knowledge, such as a notable medical issue? What if it indirectly predicts mental states like depression based on historical responses and response times?

Trust is a major factor when it comes to deploying AI in the real world. Our human society depends on trust—trust in one another, in our economic and governmental systems and in the products and services we purchase and use (https://trustworthyaibook.com). Think about the role that speed limits, seat belts and airbags play today in the auto industry compared to when the first production vehicle rolled on the highway many decades ago. Given the rapid expansion of AI, we're in a similar transition where ethical, safe, and trustworthy systems are becoming increasingly important.

Challenges

One of the most general challenges of operationalizing any AI system is understanding how the use of the model will change the dynamics in and beyond the organization, and how even the best models can produce unintended outputs (good and bad) that lead to unintended outcomes (good and bad). What will happen when current customers, prospective customers, stakeholders, and even people uninvolved with the model start to experience the model's impact, and what sorts of feedback loops will develop? Consider an insurance company that adjusts its policy based on measurements of customers' behavior. In the long term, is it significantly eroding autonomy from customers who become strongly bifurcated into groups based on risk tolerance? Will the organization lose some of its best customers and lose some of its customers most in need of its services, but be left with customers who game the system or who have nowhere else to turn?

What are the ethics involved in use case selection? In the data collection? Do the data represent historical biases? Are the data representative of everyone the model will impact? Does the organization have the right to use the data to build the model and consent from individuals whose data are involved? Can the model be edited or updated if someone later opts out of being a part of it? Is any of the data collected using dark patterns or otherwise subversively collected? Is it right to share the data across jurisdictions? Is it right to take influence from other jurisdictions on the behavior of this model? Does it erode privacy in a way that would make people uncomfortable?

Fairness and bias are other challenging and interrelated ethical aspects of operationalizing AI. There are many definitions of fairness that can be at odds with each other, and it entirely depends on the context. Should fairness be focused on making sure the proportions of outcomes are the same across races, genders, income, or other status? Or should it be based on making sure that everyone has an equal offer? What happens if a customer compares their interaction with the organization with another customer, such as when the offers are widely different? (e.g., https://www.reuters.com/article/us-goldman-sachs-apple-idUSKBN1XL038). Is the model finding proxies for historical bias and using them for unfair outcomes? For example, zip code often has a strong correlation with race and income.

Solutions

Many methodologies, software packages, and frameworks have been aimed at characterizing and addressing ethical concerns. First and foremost is clear communication with those who will be and may be affected by the outcomes of the AI model, and what is known about the model. There are many proposals and discussions regarding labeling AI models in a fashion similar to food nutrition labels. Even though none has yet been solidified as a universal standard, likely due to the diversity of AI techniques and use, communicating the design, empirical observations, and methodologies to appropriate audiences.

Use Case Selection

Before a model is built or deployed, what thought has gone into using AI/ML at all? Is the problem the sort of issue that technology can support in an ethical way? Is the inherent uncertainty or probabilistic nature of AI appropriate to the task or question? There may be aspects of a problem for which AI is a suitable input, or even a suitable decisioning agent, but it is important to consider the limits and guardrails for the technology piece of the puzzle, and how humans will interpret and internalize the technology pieces of the puzzle.

Data Provenance, Lineage and Purpose

Before an AI model is operationalized, data must be collected and curated, and the model built to the goals of the system. The most important ethical aspects in this step are to ensure that a sufficiently diverse set of people have reviewed the data and its provenance and lineage, as well as the intended outcomes of the system. The data should be sufficiently inclusive and representative of the domains and people it is intended to cover. The goals should be socialized and discussed among people who can adequately represent the desires and concerns of everyone who may be affected. And finally, it is critical that developers have adequate consents for the use and storage of data used in training and evaluating models, including how the data may be combined, and that the authors, owners, and sources of data be attributed as appropriate. Data that is used, combined, and interpreted without proper context and insights and used for purposes other than consented can yield disproportionate power, especially if the data is used against minority groups or in manipulative ways. In the context of data, this potentially malicious practice is often referred to as data colonization.

Explainability

When selecting AI algorithms, it is important to understand why the model is making a prediction. The depth of understanding of why is often described in terms of explainability and interpretability. Explainability is the ability to interrogate the model to find a justification for why a prediction or decision has been made. This explanation may or may not be how the model actually made the decision, especially if the model is a cognitively impenetrable "black box" like neural networks, but there is generally at least a good likelihood that the explanation was at least part of the reason. Well-known frameworks for explaining models include methods that tell you what inputs (features) were used to drive a decision, such as LIME, which is flexible across model types, but less reliable, and SHAP, where its reliability is partly dependent on the kind of model. Other techniques include finding counterfactuals, which show what values must be changed and to what degree in order to arrive at a different conclusion.

Interpretability

Interpretability is a stronger class of understandability than explainability, where the real reason that a model made a decision is clear, understandable, and can be independently validated outside of the model. Decision trees are an example of such methods, which are easy to understand if they are small enough. Generalized linear models, generalized additive models, and related techniques can provide good accuracy while being interpretable to a data scientist. And instance-based learning techniques, such as nearest neighbors techniques can provide the data that led to a given decision. There is a historical notion that interpretability comes at a trade-off with lowered accuracy, but techniques such as optimal sparse decision trees and some commercial offerings can provide interpretability while maintaining accuracy.

Fairness

Fairness is ensuring that the model yields approximately equal metrics for some criteria, independent of other sensitive or potentially biased fields. This might include ensuring that accuracy is consistent across genders, that older candidates are not being rejected at higher rates than younger candidates, or that data which correlates strongly with race is not negatively impacting and amplifying a historical racial bias. Fairness can be applied to many different types of errors and biases (e.g., https://en.wikipedia.org/wiki/Confusion_matrix), and there are numerous tools to help assess how fair a model is from different perspectives (e.g., https://fairlearn.org/main/quickstart.html, https://github.com/Trusted-AI/AIF360).

Correlation or Causality

The relationships between interpretability, explainability, and fairness can be complex. For example, Simpson's paradox is an infamous statistical result where the trends initially look one way. But if the data are grouped, the trends for each group turn out to be exactly the opposite. It may occur far more often in real data than people realize. Determining whether a relationship is causal or just correlation is a very difficult problem, but it can be extremely difficult to determine causality without controlled experiments, sufficient data volume, and the necessary features. When operationalizing AI, it is important to be careful when assessing correlation in the data and model's predictions with causality in the real world. Correlations may only be temporary, and changes to any aspect of the feedback loop between the community and the AI model can seemingly make new correlations or break existing ones. Further, when is it appropriate to extrapolate? Are the domains and ranges well understood? Relatedly, dealing with time series or panel data often requires specialized analysis or feature transformations.

Privacy

Ensuring that the data and model both protect users' privacy is important in many domains, even when it is not obvious. For example, purchasing data, search data, and even scores from online games can reveal information about health, mental state, preferences, identity, and activity. Privacy can matter on many levels, including for individuals, groups, and organizations. It is important to remember that even if the vast majority of the people who are affected by data do not care about privacy, that privacy is still important because there may be some whose lives may be adversely affected, and privacy ultimately can affect the incentives of everyone.

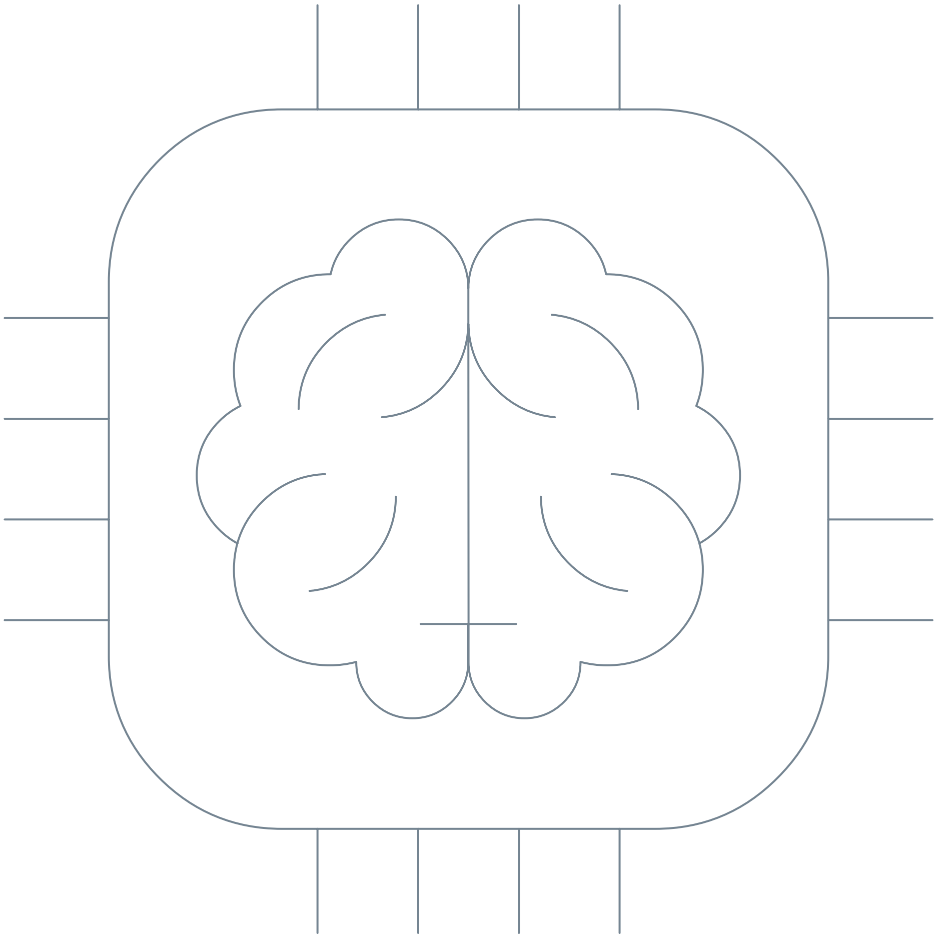
Differential privacy is one such tool that ensures there is only a marginal probability that any single individual's patterns could be discovered in the data but is best when coupled with other privacy enhancing techniques. Synthetic data is a privacy enhancing technology that ingests the original data and produces another independent data set meeting other criteria, which frequently includes maintaining all of the statistical relationships among the data while improving privacy and anonymity. And federated learning is a mechanism of training pieces of a model independently and combining them in a way that reduces memorization of individual data points. Some open-source tools exist for privacy enhancing technologies, such as differential privacy and federated learning, though the most sophisticated and easiest to use privacy enhancing technologies are currently generally commercial products.

Validation

Good data science practices are also a requirement for ethical use. The models should be checked to make sure they are not overfit and maximize the chances that the models will work well outside of laboratory conditions on historical data. Techniques such as backtesting, holding out data, or testing against new data are essential to evaluating a model, but data scientists, ML engineers, and other practitioners must be careful to make sure that they are not using their own knowledge to pollute the process. For example, flipping a coin 200 times in a row, the odds of getting 6 heads in a row is quite high. If a data scientist experiments with different methods, tunes the models extensively, and happens upon the statistical equivalent of finding those 6 heads in a row, the model could be fragile when applied to the real world. Further consideration should be applied to how bad actors may attempt to probe and exploit vulnerabilities where the model may misclassify data. Once the model has been deployed, the deployment itself may affect future data due to external knowledge of the model's use or just by changes of behavior.

Monitoring

Models should be continuously monitored, updated, investigated, and evaluated after deployment. When practical and appropriate, new data should be added, either in an online learning fashion, or to rebuild the model to keep the model relevant. Anomalous behavior, outliers, and even typical behavior should be investigated to ensure that the designs and intent of the system are still being met, are still relevant, and that those who are affected by the system directly and indirectly benefit and are not being marginalized.



Conclusion

While AI is being rapidly adopted by enterprises (large and small alike) from chatbots to recommendation engines to facial recognition to predictive analytics—the attention placed on AI ethics has been relatively low. This could be partially due to rapid advances in machine learning techniques and innovations outpacing the processes and governance required to be in place to keep the AI driven decisions in check. The consequences of not valuing AI ethics are only likely to grow. Society depends on trust. Organizations that don't value AI ethics of their data are more likely to incur reputational costs and liabilities, have trouble recruiting employees, and be left behind by competitors.

About CompTIA's Artificial Intelligence Advisory Council

CompTIA's AI Advisory Council brings together thought leaders and innovators to identify business opportunities and develop innovative content to accelerate adoption of artificial intelligence and machine learning technologies.

What We Stand For

The AI Advisory Council is committed to building the strategies and resources necessary to help companies leverage AI to be more successful. The council also collaborates with CompTIA's other industry advisory councils to further study the IT channel, blockchain, drones, business applications and internet of things. Together, we are working to:

- Discuss and explore AI business opportunities.
- Increase awareness for accelerated adoption across the tech ecosystem.
- Develop strategies to create, deliver and support AI and other emerging technologies.
- Foster greater integration of AI technology with other tech solutions.

How We're Making an Impact

As demand for data-driven analytics increases across all aspects of business, artificial intelligence is opening more doors and helping companies achieve better results, faster. The AI Advisory Council develops best practices, use cases and other resources that can be used by anyone developing, selling or influencing AI solutions.

/Administration
/Human Resources
/Legal
/Accounting
/Finance
/Marketing
/Publicity
/Promotion
/Research
/Business
/Development
/Engineering
/Manufacturing
/Planning

CompTIA®