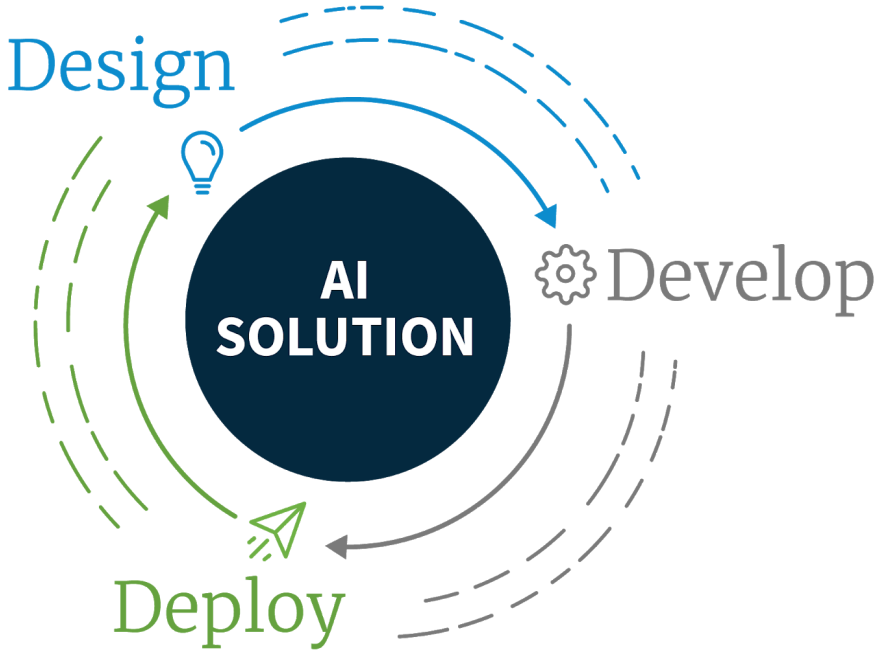
CoE Guide to AI Ethics

Artificial Intelligence (AI) is an emerging field that promises to fundamentally change the way that businesses, governments, and society as a whole interact with and impact the world around us. While some of the promises of change may be decades away, the technological possibilities are exciting. As AI continues to evolve, AI practitioners must carefully consider the impact of AI on the people who interact with it and are affected by it.

When the Centers of Excellence engage with our government partners, we promote the construction and adoption of AI applications with a human centered mindset. Development of ethical AI is good engineering and good business as AI's appropriate application in government is to improve operational efficiency and service delivery to Agency customers.

How the Centers of Excellence Use this Guide

This guide demonstrates the CoE approach to constructing AI applications with ethical considerations in mind. The questions are designed to foster discussions around broad ethics topics. Once our teams answer the questions in this guide and can defend their answers, the project moves forward. This means that each one of the questions in a given phase of the project should be answered with a "**yes**" in order to proceed to the next phase of the ethical design, development and deployment of an AI solution. These questions are continually revisited through the design-develop-deploy cycle and extensively tested to mitigate any unintended consequences from the application of AI technology.



Ethical evaluation must exist throughout the AI life cycle

Design

- Have you defined a shared goal around the outcome you are hoping to achieve with AI?
- Have you established a data governance model and a specific team responsible for ownership of data management?
- Do you have a process to establish data ownership at both a data set and individual field level?
- Have you specifically established a process to iteratively mitigate risks in the data and evaluate data for:
  - Representativeness of the population of evaluation
  - Bias towards any group(s) within that population
  - Anomalous or missing data
  - Fields that may include proprietary or personally identifying information
- Have you trained your business and technical teams about AI ethics and bias?
- Have you made a conscious effort to create a diverse and inclusive team?
- Have you established a process to assess and document the privacy implications of the AI solution?
- Have you established a process to test, monitor and evaluate the performance of your AI solution?

Develop

- Have you established a process to log the training and development process to include:
  - Alterations to the data during data wrangling and transformation
  - Which data are included in train, test and validation sets
  - Model performance on each of these data sets
- Have you established a process to identify and address ways that the algorithm could discriminate against protected classes, either by direct inclusion of these attributes or by proxy?
- Have you established a risk mitigation strategy for challenges that arise through the development lifecycle?
- Do you have a process in place to test, track and evaluate the model optimization process?
- Have you established a means to evaluate the AI solution's ability to generalize to a broader range of inputs as it moves into production?
- Have you established a transparent, participatory, and an accountable process to incorporate diverse views?
- Have you determined a mechanism to explain the model development process, including the inputs and outputs, to a non-technical audience before the AI solution is deployed?

Deploy

- Have you established a process by which you monitor the performance of your model in production?
- Have you determined quantitative evaluation criteria to measure the impact on users, stakeholders affected by the decision made or informed by the AI solution?
- Do users of the AI solution have a mechanism to see or understand how the AI system makes a decision?
- Have you established a feedback mechanism to prevent model degradation?
- Have you created a mechanism for users and stakeholders to report and explain their experience interacting with the AI solution?
- Have you established transparent, accountable, and participatory mechanisms for diverse input?
- Have you allowed a process by which an end-user can contest the outcome of the AI solution if needed?