# Artificial Intelligence Ethics Framework for the Intelligence Community v. 1.0 as of JUNE 2020

Artificial Intelligence (AI) can enhance the intelligence mission, but like other new tools, we must understand how to use this rapidly evolving technology in a way that aligns with our principles to prevent unethical outcomes. This is an ethics guide for United States Intelligence Community personnel on how to procure, design, build, use, protect, consume, and manage AI and related data. Answering these questions, in conjunction with your agency-specific procedures and practices, promotes ethical design of AI consistent with the Principles of AI Ethics for the Intelligence Community. This guide is not a checklist and some of the concepts discussed herein may not apply in all instances. Instead, this guide is a living document intended to provide stakeholders with a reasoned approach to judgment and to assist with the documentation of considerations associated with the AI lifecycle. In doing so, this guide will enable mission through an enhanced understanding of goals between AI practitioners and managers while promoting the ethical use of AI.

The use of AI must match the Intelligence Community's unique mission purposes, authorities, and responsibilities for collecting and using data and AI outputs. AI should:

- Be used when it is an appropriate means to achieve a defined purpose after evaluating the potential risks;
- Be used in a manner consistent with respect for individual rights and liberties of affected individuals, and use data obtained lawfully and consistent with legal obligations and policy requirements;
- Incorporate human judgment and accountability at appropriate stages to address risks across the lifecycle of the AI and inform decisions appropriately;
- Identify, account for, and mitigate potential undesired bias, to the greatest extent practicable without undermining its efficacy and utility;
- Be tested at a level commensurate with foreseeable risks associated with the use of the AI;
- Maintain accountability for iterations, versions, and changes made to the model;
- Document and communicate the purpose, limitation(s), and design outcomes;
- Use explainable and understandable methods, to the extent practicable, so that users, overseers, and the public, as appropriate, understand how and why the AI generated its outputs;
- Be periodically reviewed to ensure the AI continues to further its purpose and identify issues for resolution; and,
- Identify who will be accountable for the AI and its effects at each stage and across its lifecycle, including responsibility for maintaining records created.

Identifying and addressing risk is best achieved by involving appropriate stakeholders. As such, consumers, technologists, developers, mission personnel, risk management professionals, civil liberties and privacy officers, and legal counsel should utilize this framework collaboratively, each leveraging their respective experiences, perspectives, and professional skills. Agencies should also ensure that individuals involved in the design, development, review, deployment, and use of any AI have sufficient training to address the questions below and the issues presented above.

**Purpose: Understanding Goals and Risks.** Determine what goals you are trying to achieve to ensure you can design AI that balances desired results with acceptable risk.

- What is the goal you are trying to achieve by creating this AI, including components used in AI development? Is there a need to use AI to achieve this goal? Can you use other non-AI related methods to achieve this goal with lower risk? Is AI likely to be effective in achieving this goal?
- Are there specific AI system methods suitable and preferred for this use case? Does the efficiency and reliability of the AI in this particular use case justify its use for this purpose?
- What benefits and risks, including risks to civil liberties and privacy, might exist when this AI is in use? Who will benefit? Who or what will be at risk? What is the scale of each and likelihood of the risks? How can those risks be minimized and the remaining risks adequately mitigated? Do the likely negative impacts outweigh likely positive impacts?
- What performance metrics best suit the AI, such as accuracy, precision, and recall, based on risks determined by mission managers, analysts, and consumers given the potential risks; and how will the accuracy of the information be provided to each of those stakeholders? What impacts could false positive and false negative rates have on system performance, mission goals, and affected targets of the analysis?
- Have you engaged with the AI system developers, users, consumers, and other key stakeholders to ensure a common understanding of the goal of the AI and related risks of utilizing AI to achieve this goal?
- How are you documenting the goals and risks?

**Legal Obligations and Policy Considerations Governing the AI and the Data.** Partnering closely with risk management teams in your agency, including your legal, compliance, records management, classification, and civil liberties and privacy professionals, will help you understand governing authorities, legal obligations, information management responsibilities, and risks associated with an AI project.

- What authorities, agreements, or contracts govern the collection or acquisition of all sources of the data related to the model (training, testing, and operational data)? Who can clarify limitations from the agreements or contracts?
- What legal or policy restrictions exist on the use of data under this authority/agreement/contract? (For example, data subject to the Privacy Act should be used for a purpose that is compatible with that for which the data was collected).
- How must data be stored, shared, retrieved, accessed, used, retained, disseminated, and dispositioned under the authority/agreement/contract, as well as relevant constitutional, statutory, and regulatory provisions?
- What authorities or agreements apply to the AI itself, including the use, modification, storage, retrieval, access, retention, and disposition of the AI? Are there any proposed downstream applications of the AI that are legally restricted from using the underlying data?
- Does combining data with other inputs from the AI create new legal, records management, or classification risks relating to how the information is maintained and protected?

**Human Judgment and Accountability.** In the Intelligence Community, the potential purposes and applications of an AI could range from basic business tasks to highly sensitive intelligence analysis. The assessed risk will determine at which point in the process and to what degree a human will be involved in the AI.

- Given the purpose of the AI and potential consequences of its use, at what points, if any, are a human required as part of the decision process? If the AI could result in significant consequences such as an action with the potential to deprive individuals of constitutional rights or the potential to interfere with their free exercise of civil liberties, how will you ensure individual human involvement and accountability in decisions that are assisted through the use of AI?
- Where and when should the human be engaged? Before the results are used in analysis? Before the outputs are provided for follow-on uses?
- Who should be the accountable human(s)? Do they know that they are designated as the accountable human(s)? What qualifications are required to serve in that role? How is accountability transferred to another human?
- What are the access controls and training requirements for those operating at different stages in the AI lifecycle?
- What does the accountable human need to know about the AI to judge its reliability and accuracy?
- How may introducing an accountable human produce cognitive biases and/or confirmation bias?
- Who should be engaged for unresolved issues and disputes regarding the AI or its outputs?

**Mitigating Undesired Bias and Ensuring Objectivity.** Ensuring objectivity is a defining characteristic of intelligence analysis. In conducting analysis, Intelligence Community Directive 203 requires that we must perform our functions "with objectivity and with awareness of [our] own assumptions and risks. [We] must employ reasoning techniques and practical mechanisms that reveal and mitigate bias." For legal, policy, and mission reasons, however, there are certain "biases" that the Intelligence Community intentionally introduces as it designs, develops, and uses AI. Specifically, we design our models and choose our datasets to screen out irrelevant information, focus on the specific foreign intelligence targets, and appropriately minimize the collection and use of United States person information. In mitigating bias, the Intelligence Community therefore focuses on identifying and minimizing undesired bias. "Undesired bias" is bias that could undermine analytic validity and reliability, harm individuals, or impact civil liberties such as freedom from undue government intrusion on speech, religion, travel, or privacy. Undesired bias may be introduced through the process of data collection, feature extraction, curating/labeling data, model selection and development, and even in user training. Taking steps to discover bias throughout the lifecycle of an AI, mitigate undesired bias, and to document and communicate known biases and how they were addressed, are critical to long-term reliance on training data sets, to reusing models, and to trusting outputs for follow-on use.

- How complete are the data on which the AI will rely? Are they representative of the intended domain? How relevant is the training and evaluation data to the operational data and context? How does the AI avoid perpetuating historical biases and discrimination?
- What are the correct metrics to assess the AI's output? Is the margin of error one that would be deemed tolerable by those who use the AI? What is the impact of using inaccurate outputs and how well are these errors communicated to the users?

- What are the potential tradeoffs between reducing undesired bias and accuracy? To what extent can potential undesired bias be mitigated while maintaining sufficient accuracy?
- Do you know or can you learn what types of bias exist in the training data (statistical, contextual, historical, or other)? How can undesired bias be mitigated? What would happen if it is not mitigated? Is the selected testing data appropriately representative of the training data? Based on the purpose of the AI, how much and what kind of bias, if any, are you willing to accept in the data, model, and output? Is the team diverse enough in disciplinary, professional, and other perspectives to minimize any human bias?
- How will undesired bias and potential impacts of the bias, if any, be communicated to anyone who interacts with the AI and output data?

**Testing Your AI.** Every system must be tested for accuracy in an environment that controls for known and reasonably foreseeable risks prior to being deployed.

- Based on the purpose of the AI and potential risks, what level of objective performance for the desired performance metric, (e.g., precision, recall, accuracy, etc.) do you require?
- Has the AI been evaluated for potential biased outcomes or if outcomes cause an inappropriate feedback loop? Have you considered applicable methods to make the AI more robust to adversarial attacks?
- How and where will you document the test methodology, results, and changes made based on the test results?
- If a third party created the AI, what additional risks may be associated with that third party's assumptions, motives, and methodologies? What limitations might arise from that third party claiming its methodology is proprietary? What information should you require as part of the acquisition of the analytic? What is the minimum amount of information you must have to approve an AI for use?
- Was the AI tested for potential security threats in any/all levels of its stack (e.g. software level, AI framework level, model level, etc.)? Were resulting risks mitigated?

**Accounting for Builds, Versions, and Evolutions of an AI.** A successful AI is often refined through numerous iterations, versions, or evolutions, both while it is being trained on training data and after it matures and is applied to mission, analytic, and business data. An existing AI may also be repurposed and require modifications and/or retraining prior to redeployment.

- As you refine the AI, how does the data you have used, the parameters and weights you have chosen, and the outputs ensure that this version or evolution is designed to achieve the authorized purpose?
- Have you accounted for natural data drift within the operational environment compared to training data?

- Have you documented provenance of data, outputs of the iteration, and test results (accuracy) in a way that will provide for repeatability, auditing, and oversight? If the AI is continuously modified, are all critical aspects, dependencies, and artifacts version controlled and documented? Will it be clear to anyone auditing the AI or consumers of the AI's outputs which version was in use at any given moment in time? Will it be clear which iteration of a model drew on which data and produced what outputs?
- Where will you save documentation on versions of AI and relevant training and test data? Have you made that information available to users and consumers of the AI? How will this documentation be retained and made discoverable to ensure compliance with your Agency's records management responsibilities?
- Have you accounted for changes in demographics of your customer for your AI capability (e.g., changing user experience needs) or the changing needs of the mission?

**Documentation of Purpose, Parameters, Limitations, and Design Outcomes.** To ensure your AI is used properly, it is important to communicate (a) what the AI is for, (b) what it is not for, (c) how it was designed, and (d) what its limitations are. Documentation assists not only with proper management of the AI, but also with determining whether the AI is appropriate for new purposes that were not originally envisioned.

- How can you store the documentation in a way that is available to all potential consumers of this AI?
- Have you documented where the data came from and its downstream uses and sharability? The downstream uses and sharability of the AI?
- Have you documented what rules apply to the data as a whole? What rules apply to subsets?
- Have you documented the potential risks of using the AI and its output data, and the steps taken to minimize these risks?
- Have you documented use cases for which the AI was and was not specifically designed?
- Have you documented the process for discovering undesired bias and the conclusions?
- Have you documented how to verify and validate the model as well as the frequency with which these checks should be performed?

**Transparency: Explainability and Interpretability.** AI can achieve the anticipated outcome despite using inappropriate criteria. Consistent with the Principles of Intelligence Transparency for the Intelligence Community, use methods that are explainable and understandable, to the extent practicable, so that users, overseers, and the public, as appropriate, understand how and why the AI generated its outputs.

- Given the purpose of the AI, what level of explainability or interpretability is required for how the AI made its determination? If a third party created the AI, how will you ensure a level of explainability or interpretability? Does this conform with Intelligence Community Directive 203: Analytic Standards?
- How are outputs marked to clearly show that they came from an AI?
- How might you respond to an intelligence consumer asking "How do you know this?" How will you describe the dataset(s) and tools used to make the output? How was the accuracy or appropriate performance metrics assessed? How were the results independently verified? Have you documented and explained that machine errors may differ from human errors?

**Periodic Review.** All AI should be checked at an appropriate documented interval to determine whether it still meets its purpose and that any undesired biases or unintended outcomes are appropriately mitigated.

- How will user and peer engagement be integrated into the model development process and periodic performance review once deployed?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- As time passes and conditions change, is the training data still representative of the operational environment?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed? How much distributional shift or model drift from baseline performance is acceptable?
- Who is responsible for checking the AI at these intervals?
- How will the accountable human(s) address changes in accuracy and precision due to either an adversary's attempts to disrupt the AI or unrelated changes in the operational/business environment, which may impact the accuracy of the AI?

**Stewardship and Accountability: Training Data, Algorithms, Models, Outputs of the Models, Documentation.** Before the AI is deployed, it must be clear who will have the responsibility for the continued maintenance, monitoring, updating, and decommissioning of the AI.

- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
- Who is accountable for the ethical considerations during all stages of the AI lifecycle?
- If anyone believed that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?