Ethical AI Guidelines & Processes

Last updated by | David Cox | May 20, 2024 at 11:23 AM PDT

PURPOSE: The Rethink AI Ethical Checklist is meant as a documentation guide and prompt to ensure that developers of AI products at Rethink consider ethical practices of AI development. It is adapted from the NeurIPS Code of Ethics.

DATA-RELATED CONCERNS: The points listed below apply to all datasets used to develop AI products.

Item	Relevant to your product?	Notes on relevance and how addressed (user encouraged to attach links).
Privacy: Have you minimized the use and exposure of any personally identifiable information (PII), personal health information (PHI), and student education records (SER)?		
Legal Use of Data: Have you confirmed your use of the data complies with the end user license agreement from the product from which that data were collected?		
Deprecated datasets: Have you documented the statistical distributions of your data and established boundary conditions for useful and appropriate uses of the model?		
Representative evaluation practice: Have you assessed and documented how well the data used to build the AI product aligns with the characteristics of the users who will use the product?		
Tracking Model Drift and Degradation: Have you established data pipelines to monitor, log, and report input drift and change in loss metrics from development models?		

SOCIETAL IMPACT & POTENTIAL HARMFUL CONSEQUENCES: Developers should transparently communicate the known or anticipated consequences of the product use. The following specific areas are of particular concern:

Item	Is this relevant to your product?	Notes on relevance and how addressed.
Safety: Are there foreseeable situations in which the technology can cause harm or injury through its direct application, side effects, or potential misuse?		
Security: Is a risk that the applications could open security vulnerabilities or cause serious accidents when deployed in real world environments?		
Discrimination: Can the technology developed be used to discriminate, exclude, or otherwise negatively impact people, including impacts on the provision of services such as healthcare or education?		
Bias and fairness: Have you assessed and documented any potential biases or limitations to the scope of performance of models or the contents of datasets? For example, have you inspected these to ascertain whether they encode, contain or exacerbate bias against people of a certain gender, race, sexuality, or other protected characteristics.		

IMPACT MITIGATION MEASURES: It is important to reflect and take action to mitigate any potential harmful consequences that may result from an AI product.

Item	Is this relevant to your product?	Notes on relevance and how addressed.
Data and model documentation: Have you communicated the details of the dataset or the model via a structured template?		
Data and model artifacts: If releasing data or models for others at Rethink to use, have you documented the intended use and limitations of these artifacts to prevent misuse or inappropriate use?		
Secure and privacy-preserving data storage & distribution: Have you adhered to Rethink standards around privacy protocols, encryption and anonymization to reduce the risk of data leakage or theft?		
Responsible release and publication strategy: If your model has a high risk for misuse or dual-use, have you released it with the necessary safeguards to allow for controlled use of the model? For example, by requiring that users adhere to a code of conduct to access the model.		
Allowing access to research artifacts: Have you made accessible the information required to enable scrutiny and auditing (e.g., information required to understand your code, execution environment versions, weights, hyperparameters of systems, etc.)? This should be accomplished in a manner allowing for the sufficient reproduction of described results.		