Working Draft Artificial Intelligence Operating Principles Under Development

(Presented by Alka Roy, AT&T, on behalf of AT&T to the *Trusted AI Committee* of the *LF AI Foundation*)

November 7, 2019 Version

ABSTRACT

Earlier this year, AT&T released our AI Guiding Principles. Since then, throughout the many parts of AT&T, we continuing to develop, test and enhance our practices to operationalize trusted AI. There is no one-size-fits-all solution out there to be found for all of AT&T's operations, and as an industry and globally we will continue to discover new areas and ways of realizing AI trustworthiness, so we see the develop/test/enhance cycle as ongoing. On the other hand, some important and broadly applicable basic methods and principles have risen to the top consistently across a number of contexts. At AT&T, we have compiled a working draft set of such methods and principles gathered from our perspective and experience so far in order to share it with the LF AI committee as a seed. We hope that this contribution will spur the discussion and enquiries in this area as well as this community's goal to develop consensus among representative and consequential stakeholders in support of a basic set of Trustworthy AI Operating Principles generally applicable to and effective across a broad array of AI activities in a practical, efficient and operational manner.

Operating Pillar	Operating Principle	Explanation
#1 Risk Assessment and Planning[1]	Bias	Identify AI Solutions[2] that could trigger legal, policy or ethical concerns resulting from unintended bias in the learning or operational data sets or in the properties of the rules or algorithm (e.g., unlawful discrimination against a protected class, solutions that pose difficulties for those with disabilities, or use of data elements that effectively serve as a proxy for race or ethnicity).
	Safety	Identify AI Solutions that have the potential to diminish physical safety.
	Privacy[3]	Identify AI Solutions that may adversely affect privacy interests, including in data collection and usage and AI Solution outputs.
	Accessibility	Identify AI Solutions that have the potential to pose difficulties for those with disabilities, including in business to consumer applications of AI, in which a consumer interacts directly with an AI Solution with no human intermediary.
	Law	Identify AI Solutions that affect rights under domain-specific laws (e.g., EEO, FCRA) or AI-specific laws and regulations not otherwise addressed by prior principles.
	Consequential Use Cases	Identify AI Solutions that provide access to services which, if denied or diminished, could have meaningful and significant impact on an individual in areas of employment, communication and entertainment (e.g., digital divide/redlining, PII/CPNI, network prioritization, national security) to the extent not otherwise addressed by prior principles.

Operating Pillar	Operating Principle		Explanation
#2 Data Governance[4]	Data Propert.	Data Provenance	Create and maintain a human readable and accessible record of practices and actions taken on data used in the AI Solution to enable data professionals to understand the lineage of the data used throughout the AI lifecycle.

[1] This Pillar should be used before the others to identify AI solutions that present higher risks. AI Solutions that present higher risk potential should be accorded more resources for implementing some or all (as appropriate to the risk profile of the AI solution) of the balance of the Pillars and Operating Principles to prevent, mitigate and remediate risk.
[2] "AI Solution" means an application of one or more "AI Systems" to solve a human-defined problem or achieve a human-defined goal. "AI System" means a machine-based system that can, for a given set of human-defined objectives, train on data and develop models to make predictions, recommendations, or decisions influencing real or virtual environments. These data-driven AI systems are designed to operate with varying levels of autonomy.
[3] In many enterprises, there will be an extant Privacy assurance infrastructure that can be applied to AI operations.
[4] In many enterprises, there will be an extant Data Governance infrastructure that can be applied to AI operations.

Operating Pillar	Operating Principle	Explanation
	Data Quality	Develop and implement controls to identify and eliminate unintended biases, inaccuracies, errors and mistakes in the data used throughout the AI lifecycle.
	Data Integrity	Develop and implement controls to prevent the insertion of malicious data or corruption of data.
	Third Party Data	When using third party data, select from a diversity of appropriate sources, and assess such data for compliance with applicable AI Operating Principles and policies.
	Data Access Management	Limit (and routinely audit) data access privileges to those with appropriate level of competence and a legitimate business need.

Operating Pillar	Operating Principle		Explanation
#3 Prevention and Minimization of Negative Impacts (NOS)[5]	Human Properties	Human Direction	Be clear and document the purpose for an AI Solution.
		Human Oversight	Subject AI Solutions to appropriate and relevant human oversight for accountability, biases, performance, inaccuracies and other defects.
		Team Diversity	Encourage diversity for personnel working on AI Solutions that is consistent with HR policies and values.
		Ethical AI Training	Using best practices, develop and require training of personnel working on AI Solutions on responsible and trustworthy AI.
		Stakeholder Participation	Consult a diverse sample of the individuals affected by the AI Solution throughout the lifecycle where feasible.
	Model Properties	Accuracy	Ensure the AI Solution's results (predictions, recommendations or decisions) accurately reflect the human-defined purpose for the AI Solution.
		Reproducibility	Ensure that the AI Solution produces the same results when repeated under the same conditions.
		Reliability	Ensure the AI Solution produces accurate and acceptable results across a range of inputs and situations.
		Tools for Trustworthy AI	Employ available, effective technological tools designed to mitigate biases and inaccurate outcomes as well as operate within Trustworthy AI principles at appropriate stages of the AI lifecycle.
		Third Party AI/ML Providers	Assess third party AI/ML providers vis a vis AI Operating Principles and policies. Where feasible, assess and audit for compliance with applicable AI Operating Principles and policies. Develop relevant contract language for vendor agreements.
	Contingency Planning	Security and Resiliency	Develop the AI Solution to mitigate/prevent unintended manipulation and/or exploitation by adversaries including through adversarial testing where warranted. Build best-practice to stay current with security and resiliency methodologies and build process to implement as needed.
		Fallback Planning	Design the AI Solution to include a transparent and identifiable fallback plan if problems are uncovered (e.g., switching from a learning to a rules-based model, assertion of human control).

[5] This Pillar contains principles for preventing and minimizing risk that are not otherwise specified in Pillars 2 and 4.

	Change Management	Reassess the risk and adjust risk mitigation and remediation tactics as appropriate when an AI System or AI Solution is re-tasked to solve a new problem or when there are changes in input data, stakeholder composition or other material conditions. Document per policies.
	Collaboration and Sharing	Actively participate in industry-wide fora (including open source) designed for AI actors to collaborate on and open standards, share tools, methods and models for trustworthy AI. Continue to adapt and evolve specific tools and methodologies across the company.

Operating Pillar		Operating Principle	Explanation
#4 Accountability	Internal	Roles and Responsibilities	Clearly define the roles and responsibilities of team members selling, designing, developing, building and maintaining AI Solutions.
		Documentation	Record each material step, decision and rationale in the AI lifecycle, as well as outputs and outcomes relative to the original objective as well as these AI operating principles.
		Traceability	To the extent technically feasible, identify and clearly document the inputs and processes that affect the output of the AI Solution for later identification of the reasons for outputs.
		Auditing	Audit AI Solutions, with the frequency and depth of the audit reflecting the underlying risk of the AI Solution, engaging independent third-party audits for higher risk AI Solutions.
	External	Transparency	Provide public visibility to relevant, general information in language that is readily understandable to the average adult member of the public that summarizes how the Enterprise uses AI Solutions.
		Disclosure	Where appropriate, affirmatively disclose - in terms that are understandable to persons interacting with the AI Solution - the existence of an AI Solution with other information appropriate under the circumstances (which might include identifying the types of data used by the AI Solution or explaining how the Enterprise is using the AI Solution).
		Explainability	To the extent technically feasible and relevant, explain the outputs of an AI Solution in terms understandable to the requester.
		Opportunity to Challenge and seek Redress	Provide persons for whom an AI Solution's outputs has produced a materially adverse effect an opportunity to challenge the output and seek redress.