Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-risked, respectful, secure, honest, and usable artificial intelligence (AI) systems with a diverse team aligned on shared ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development* by Carol Smith, available at https://arxiv.org/abs/1910.03515.

We will design our AI system with the following in mind:

- ☐ Designated humans have the ultimate responsibility for all decisions and outcomes:
  - Responsibilities are explicitly defined between the AI system and human(s), and how they are shared.
  - Human responsibility will be preserved for final decisions that affect a person's life, quality of life, health, or reputation.
  - Humans are always able to monitor, control, and deactivate systems.
- ☐ Significant decisions made by the AI system will be
  - explained
  - able to be overridden
  - appealable and reversible

We work to speculatively identify the full range of risks and benefits:

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI system, including the following:

- ☐ communication plans to share pertinent information with all affected people
- ☐ mitigation plans for managing the identified speculative risks

We value respect and security:

- ☐ incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity, and inclusion
- ☐ respecting privacy and data rights (Only necessary data will be collected.)
- ☐ providing understandable security methods
- ☐ making the AI system robust, valid, and reliable

We value transparency with the goal of engendering trust:

- ☐ The purpose, limitations, and biases of the AI system are explained in plain language.
- ☐ Data sources have unambiguous respected sources, and biases are known and explicitly stated.
- ☐ Algorithms and models are appropriate and verifiable.
- ☐ Confidence and context are presented for humans to base decisions on.
- ☐ Transparent justification for recommendations and outcomes is provided.
- ☐ Straightforward and interpretable monitoring systems are provided.

We value honesty and usability:

- ☐ Humans can easily discern when they are interacting with the AI system vs. a human.
- ☐ Humans can easily discern when and why the AI system is taking action and/or making decisions.
- ☐ Improvements will be made regularly to meet human needs and technical standards.

Team Signatures and Date

About the SEI

The Software Engineering Institute is a federally funded research and development center (FFRDC) that works with defense and government organizations, industry, and academia to advance the state of the art in software engineering and cybersecurity to benefit the public interest. Part of Carnegie Mellon University, the SEI is a national resource in pioneering emerging technologies, cybersecurity, software acquisition, and software lifecycle assurance.

Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612

sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu