Responsible AI
DoD's Ethical Principles for AI

Governable

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM22-0009

Governable

Governable

Design and engineer AI capabilities to fulfill intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Brakes, back doors, and buffers

Intentional design – Be thoughtful and thorough in planning and development.

Make AI system robust and secure, valid, and reliable.

Be speculative – conduct curiosity activities from user experience and human-computer interaction practices.

Respectful and secure

Respect privacy and data rights.

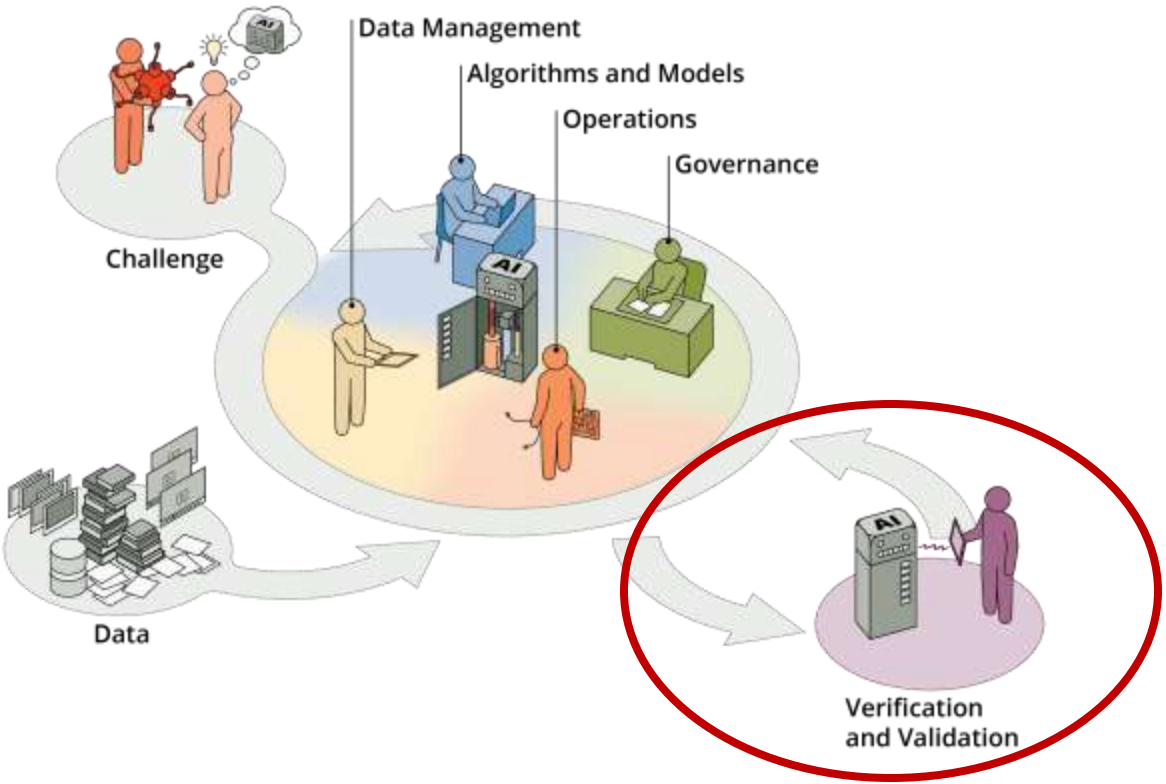
“Collect only information that is reasonably necessary to conduct its mission in a lawful manner.” –Defense Innovation Board

Provide understandable security.

Show awareness of known and desirable bias.

Provide ways to report undesirable bias and issues.

Constant verification and validation



Reward team members for finding ethics bugs

Dr. Ayanna Howard
- on the Artificial Intelligence Podcast with Lex Fridman

Mitigation and communication plans

- Who can report?
- To whom?
- What is the method for "turning it off"?
- Who has access?
- What are the consequences?
- Who will be notified?
- What is the backup plan for the system?

What can you measure to track your progress?

Team

- Proactiveness in addressing bias
- More quality conversations vs. unpleasant surprises

End User

- More evidence-based decisions
- Less reports of bias

Governable – Review

Design and engineer AI capabilities to fulfill intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

Make AI respectful and secure.

Constantly verify and validate.

Make mitigation and communication plans.

Track your progress.

Empower diverse teams in inclusive environments.

Encourage deep conversations, speculation, and imaginative thinking.

Alex Van Deusen
Design Researcher
arvandeusen@sei.cmu.edu

Carol Smith
Sr. Research Scientist
cjsmith@sei.cmu.edu

Rachel Dzombak
Digital Transformation Lead
rdzombak@sei.cmu.edu