

Responsible AI

DoD's Ethical Principles for AI

Governable

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Document Markings

Copyright 2022 Carnegie Mellon University.

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.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

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.

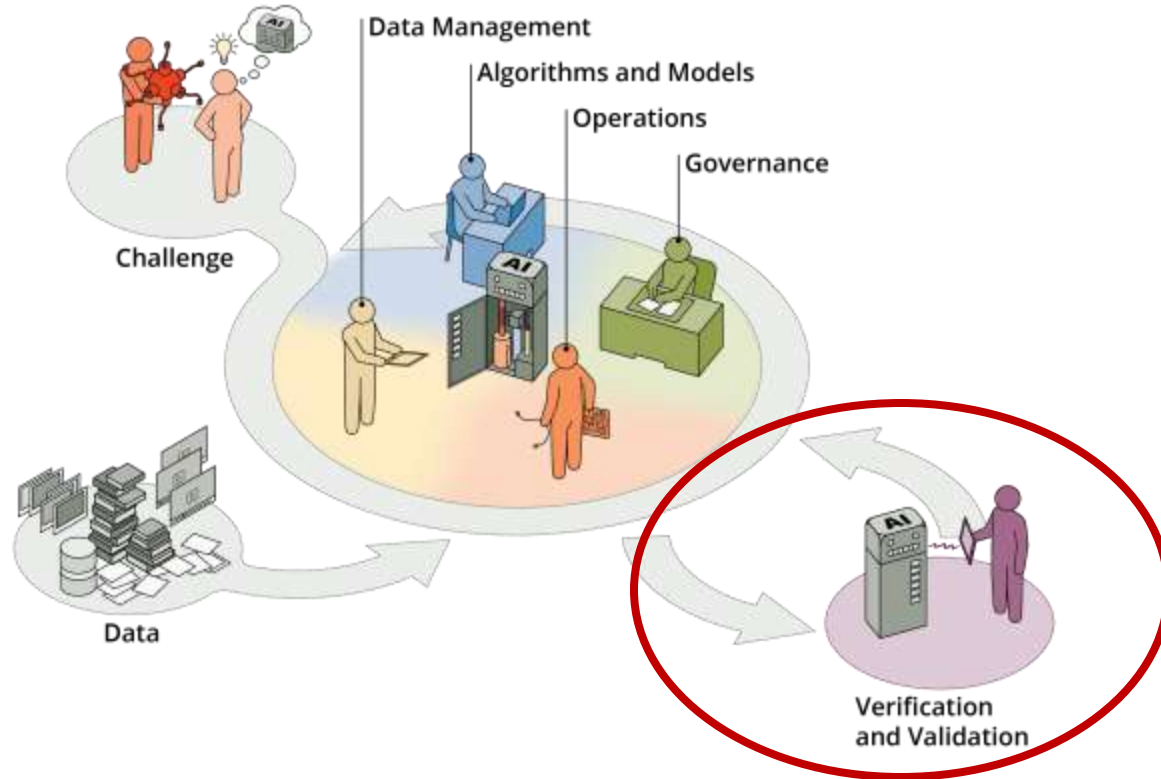
“Collect only information that is reasonably necessary to conduct its mission in a lawful manner.” –Defense Innovation Board

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