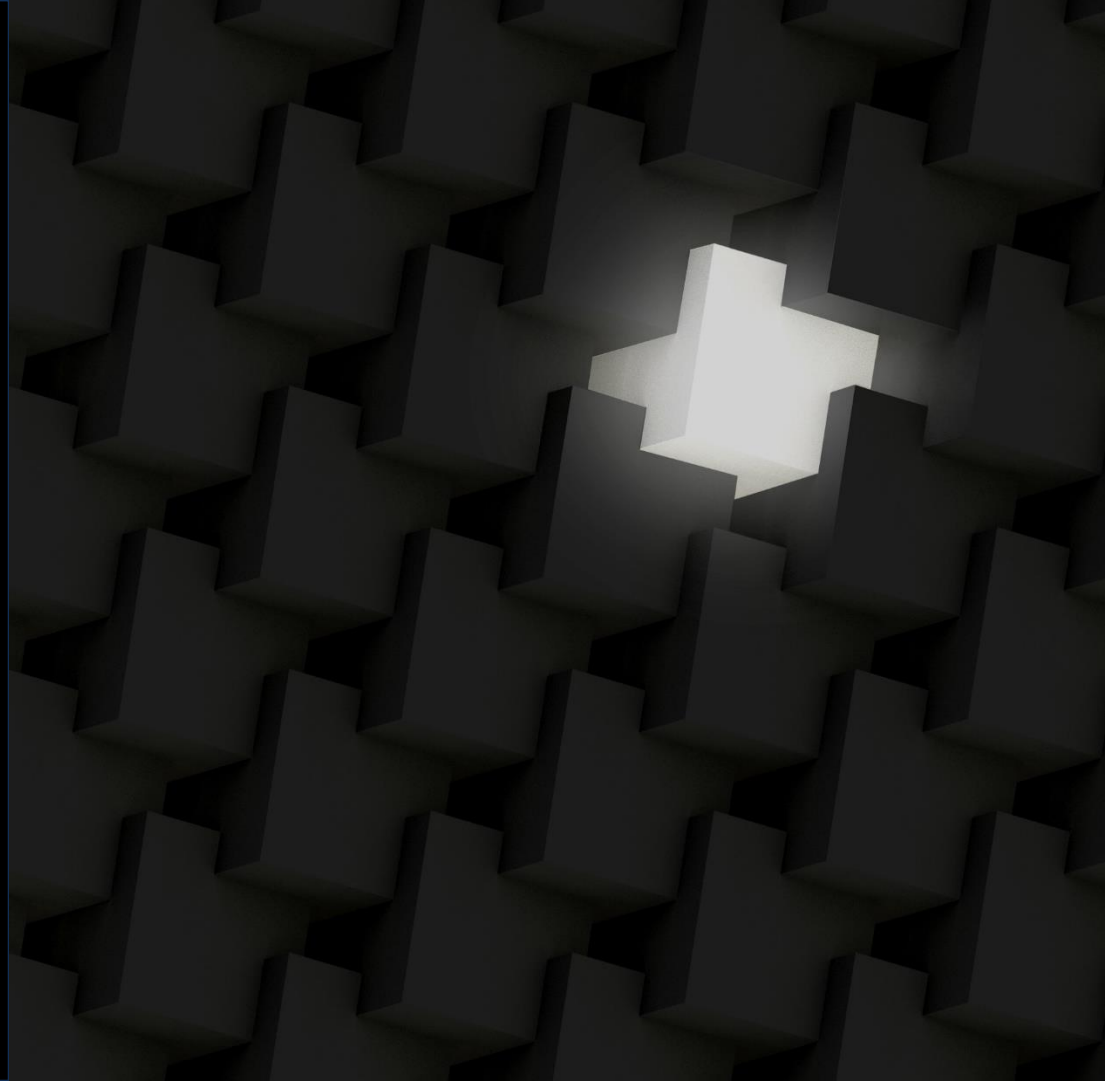


Carnegie Mellon University
Software Engineering Institute

RESEARCH REVIEW 2020

Ethics in AI Engineering

Carol J. Smith



Document Markings

Copyright 2020 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.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

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.

DM20-0929

RESEARCH REVIEW 2020

Ethics in AI Engineering

Background

Ethics

Ethics are based on well-founded standards of right and wrong.

They are the standard of expected behavior that guides the correct course of action.

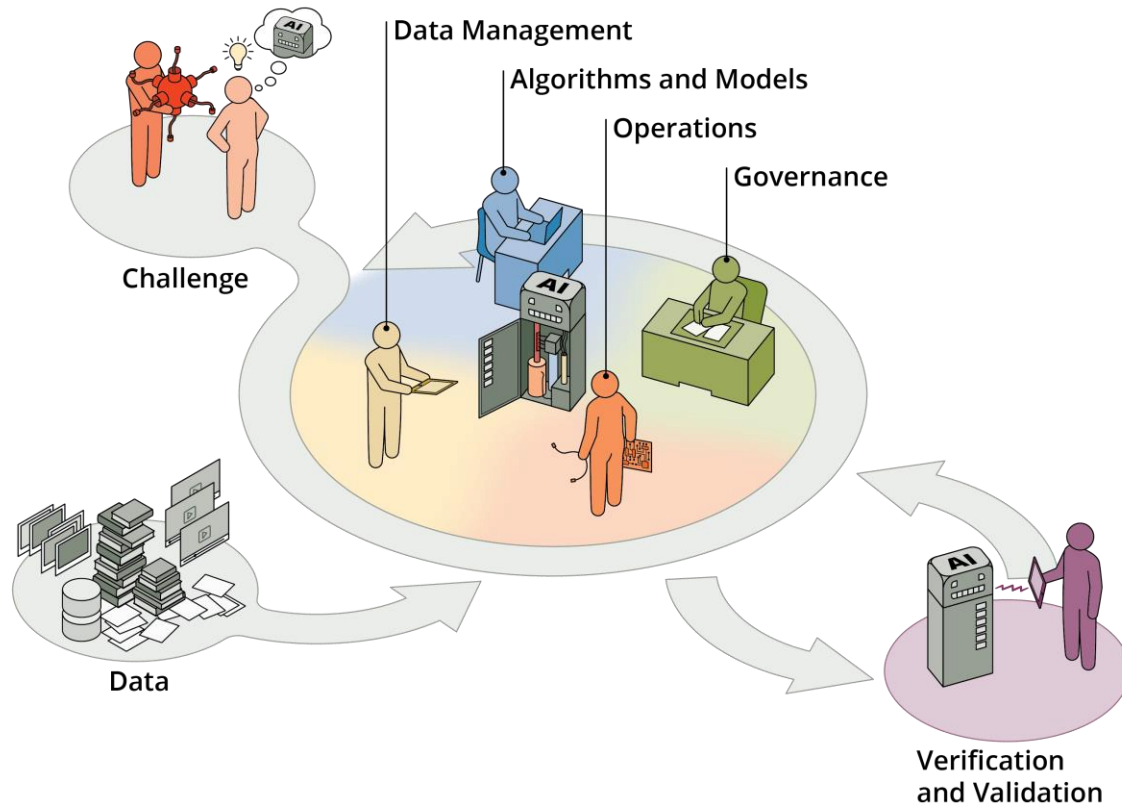
- Markkula Center for Applied Ethics, Santa Clara University

“Ethical considerations are an inseparable part of research, design, and deployment for DOD AI systems.” - Defense Innovation Board

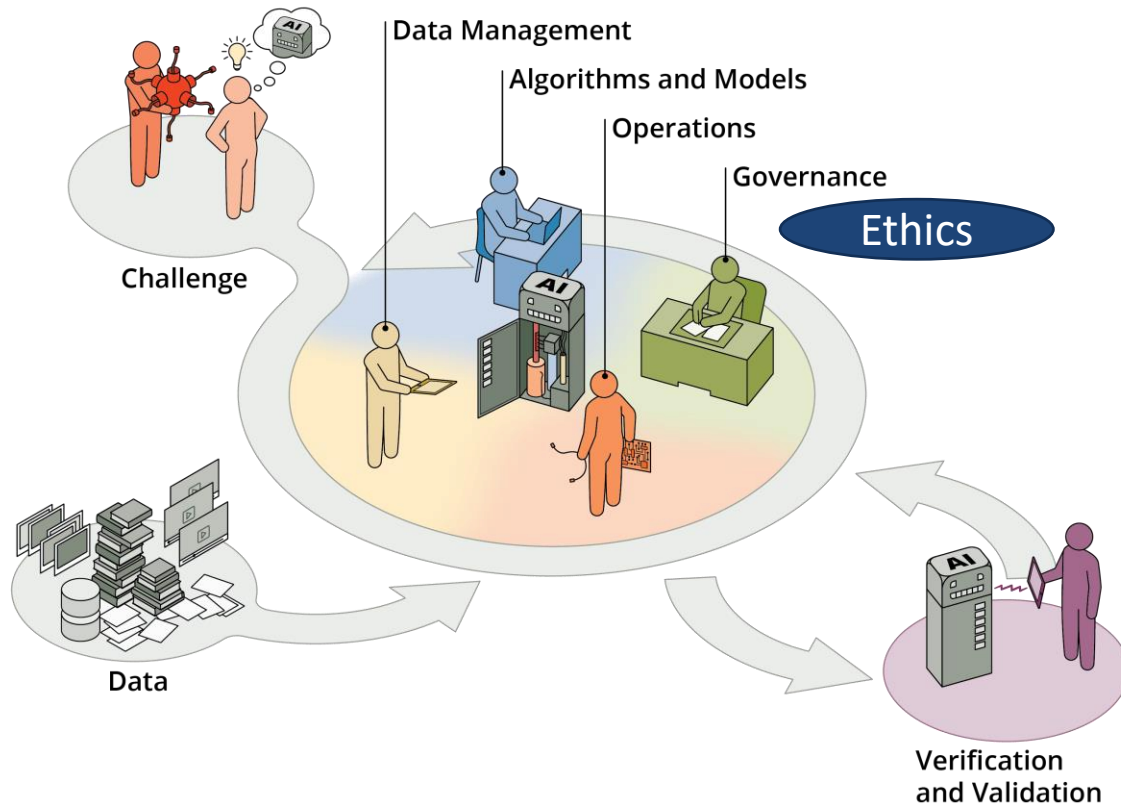
Manuel Velasquez, Claire Andre, Thomas Shanks, S.J., and Michael J. Meyer. What is Ethics? Markkula Center for Applied Ethics, Santa Clara University. <https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/>

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. Supporting Document. Defense Innovation Board. 2019.

SEI AI Engineering Process



SEI AI Engineering Process



AI has great potential; develop with caution.

Many years from now, AI may be a trustworthy substitute for human cognition and abilities.

However, for moral and legal purposes, humans must continue to deliberate over situations that involve a person's

- life (the use of force)
- quality of life
- health
- reputation

“AI will ensure appropriate human judgement and not replace it”
– Defense Innovation Board

Smith, Carol (2020): Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. Carnegie Mellon University. Conference contribution. <https://doi.org/10.1184/R1/12119847.v1>

Quote from: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. Supporting Document. Defense Innovation Board. 2019.

Why are ethical principles for AI important?

Ethical principles improve consistency:

- The team coalesces on shared set of technology ethics.
- The principles harmonize cultural variations across diverse teams.

Ethical principles aid in achieving the following:

- build on existing ethics – specific for technology
- balance to pace of change, industry pressure
- provide expectation and explicit permission, to consider and question breadth of implications

Smith, Carol (2020): Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. Carnegie Mellon University. Conference contribution. <https://doi.org/10.1184/R1/12119847.v1>

Who are technical ethics for?

Everyone creating an AI system should use technical ethics to ensure the work remains human-centered:

- data scientists and data creators
- product managers
- machine learning experts
- programmers, system architects
- curiosity experts
 - focus on situation, user's abilities, and context
 - activate curiosity via user experience (UX) activities
 - include titles such as UX researchers, human-computer/machine interaction practitioners, and digital anthropologists

Early, purposeful work

Ethics efforts are not just about the algorithms.

They're about conduct, interactions, intents and maintaining trust.

Effort to reduce risk and unwanted bias

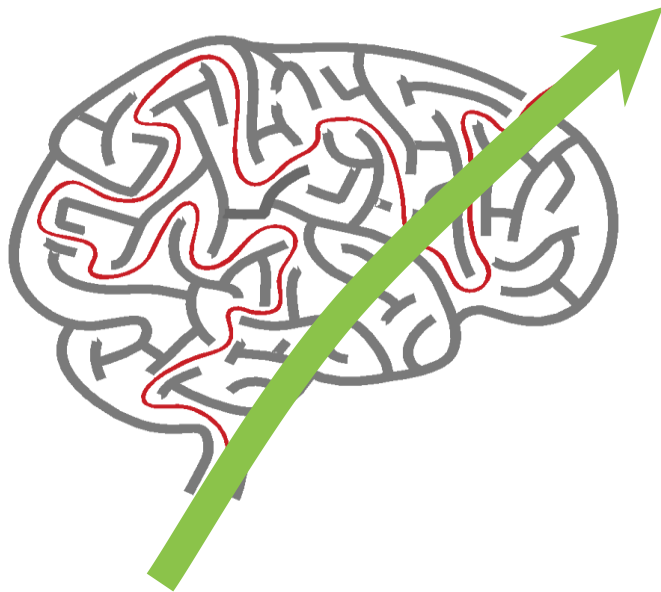
Consider benefits and harms to individuals.

- Examine initially and over time. (Are there changes?)
- Can harm be mitigated? Are mitigations a tradeoff against a benefit?

Speculate about misuse and abuse.

- Focus on potential severe abuse and consequences.
- Understand concerns/fears of those affected by AI system to identify potential misuse and abuse.

To be biased is to be human.



Biases are shortcuts, to avoid risk and simplify problems.

- not inherently bad, may be misapplied
- implicit = invisible
- not necessarily in sync with our conscious beliefs

Biases can be managed and changed.

Talk about biases in non-threatening, productive ways.

Bias is intrinsic to our experiences.

Bias can emanate from our

- social class
- resource availability
- education
- race, gender, sexuality
- culture, theology, tradition
- more...

“We often have no way of knowing when and why people are biased.”

- Sandra Wachter,
Associate Professor and Senior
Research Fellow,
University of Oxford.
Fellow, The Alan Turing Institute.

All systems have some form of bias

Complete objectivity is misleading.

Bias can have purpose and be helpful.

Our goals must be to

- reduce unintended and/or harmful bias
- prevent the inevitable harm that comes with “unknowable” systems

Bias enters via...

Data source

- all data created by humans
- curated and organized based on bias of researcher/organization

Algorithms

- sentencing and crime prediction – racial bias
- facial recognition – race and gender bias

Training

- purposeful – towards a particular outcome
- inadvertent due to content

New uncomfortable work

“Be uncomfortable.”

- Laura Kalbag,
Author, Designer.

Ethical design is not superficial.

Diverse, talented and multi-disciplinary teams

Bringing their varied skills sets, problem framing approaches, and knowledge together.

Diversity has many components:

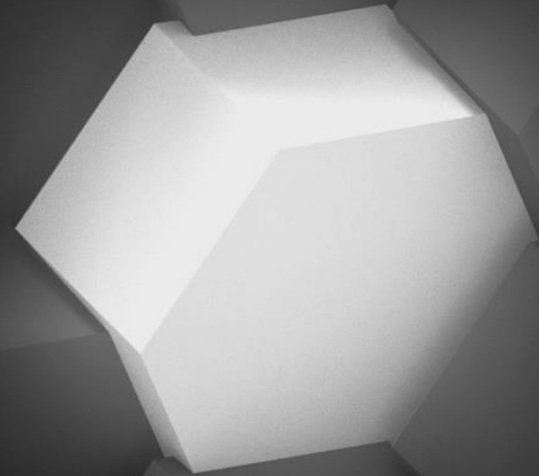
- gender, race, culture
- education (school, program, etc.)
- experiences
- thinking processes
- age, disability, and health status and more...



Great minds think differently

Not lowering the bar

Extending it



Diverse teams mitigate bias

Diverse teams

- focus more on facts
- process facts more carefully
- are more innovative

“...become more aware of their own potential biases — entrenched ways of thinking that can otherwise blind them to key information and even lead them to make errors in decision-making processes.”

David Rock, Heidi Grant. 2019. Why Diverse Teams Are Smarter. *Harvard Business Review*. November 4, 2019. <https://hbr.org/2016/11/why-diverse-teams-are-smarter>

RESEARCH REVIEW 2020

Ethics in AI Engineering

Grand Challenge: Making Ethical AI

Release

IMMEDIATE RELEASE

DOD Adopts Ethical Principles for Artificial Intelligence

FEB. 24, 2020



The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations provided to Secretary of Defense Dr. Mark T. Esper by the Defense Innovation Board last October.

<https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>

How do we make AI ethical?

Ethical Principles

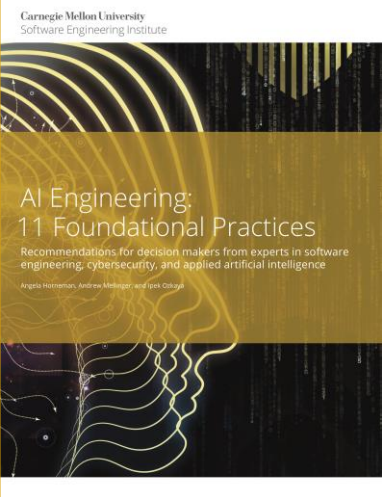
- Responsible
- Equitable
- Traceable
- Reliable
- Governable



Trustable,
Ethical AI

Content adapted from Department of Defense Adopts Ethical Principles for Artificial Intelligence - Feb. 24, 2020
<https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>

SEI: 11 Foundational Practices



Available for Download Today

AI Engineering:
11 Foundational Practices

“Developing viable and trusted AI systems that are deployed to the field and can be expanded and evolved for decades requires significant planning and ongoing resource commitment.”

Download Today

Authors

Angela Horneman, Analysis Team Lead
Carnegie Mellon University Software Engineering Institute

Andrew Mellinger, Sr., Software Developer
Carnegie Mellon University Software Engineering Institute

Ipek Ozkaya, Principal Researcher
Carnegie Mellon University Software Engineering Institute

For more information, write to sei@sei.cmu.edu

Treat ethics as both a software design consideration and a policy concern.

Incorporate user experience (UX) and interaction to constantly validate and evolve models and architecture.

Design for the interpretation of the inherent ambiguity in the output.

Horneman, Angela; Mellinger, Andrew; and Ozkaya, Ipek. AI Engineering: 11 Foundational Practices for Decision Makers.
https://insights.sei.cmu.edu/sei_blog/2019/12/ai-engineering-11-foundational-practices-for-decision-makers.html

SEI: Checklist and Agreement

Prompt conversations for understanding.

Test application of ethics.

Reduce risk and unwanted bias.

Support inspection and mitigation planning.



Checklist and Agreement - Downloadable PDF at SEI:
<https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620>

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
 - appealing 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

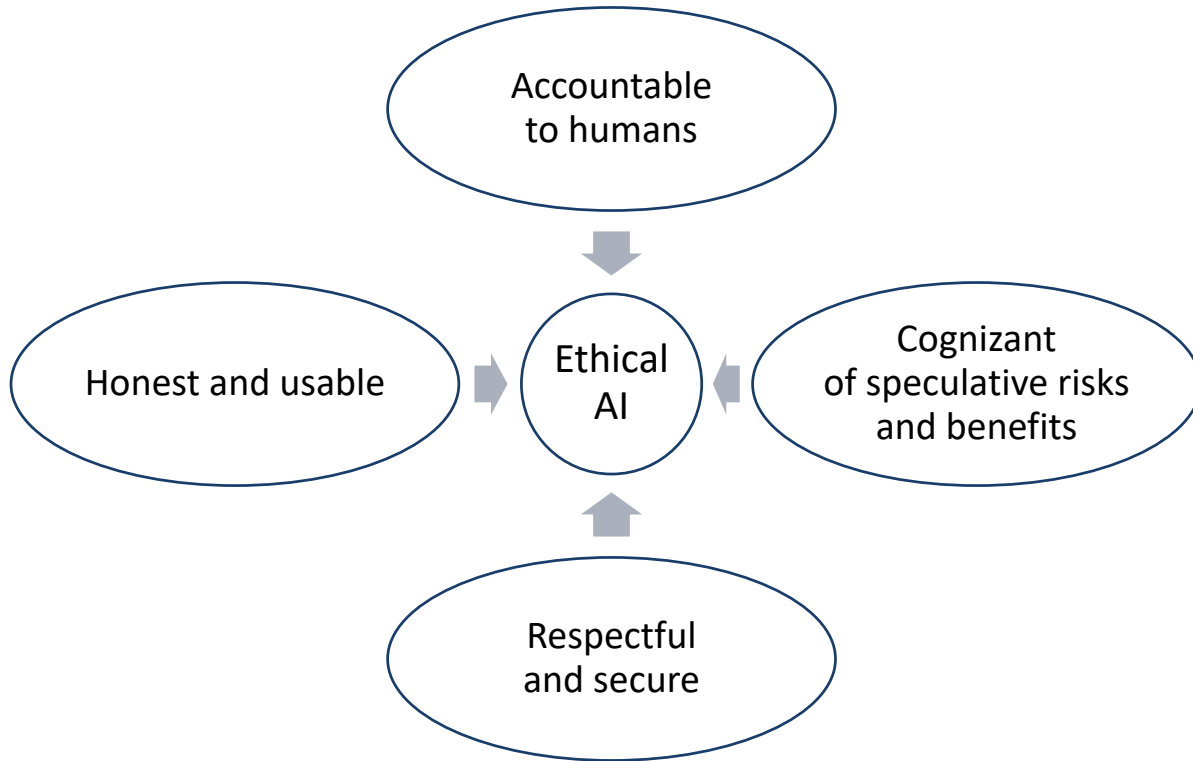
About the SEI

The Software Engineering Institute is a federally funded research and development center (FFROC) 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

Framework for Designing Trustworthy AI



Smith, Carol (2020): Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. Carnegie Mellon University. Conference contribution. <https://doi.org/10.1184/R1/12119847.v1>

Prompts help reveal hidden tasks.

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 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.



Checklist and Agreement - Downloadable PDF at SEI:
<https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620>

Create communication and mitigation plans.

Plan for unwanted consequences.

Misuse and abuse of AI system

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

Conversations for understanding

Difficult Topics

- What do we value?
- Who could be hurt?
- What lines won't our AI cross?
- How are we shifting power?*
- How will we track our progress?

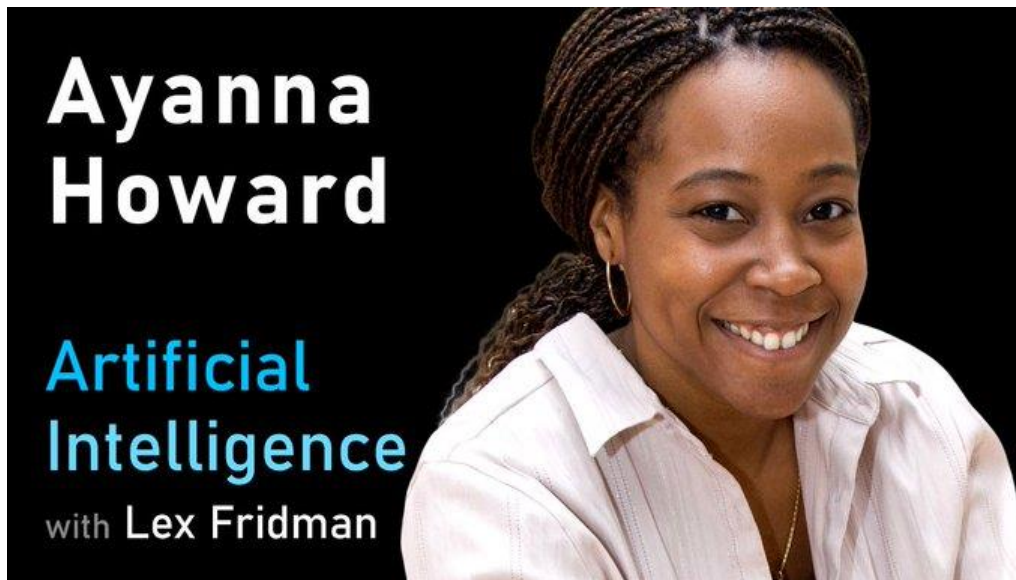
*"Don't ask if artificial intelligence is good or fair, ask how it shifts power." Pratyusha Kalluri.

<https://www.nature.com/articles/d41586-020-02003-2>

Reward team members for finding ethics bugs.

Dr. Ayanna Howard

- on the Artificial Intelligence Podcast with Lex Fridman



Dr. Ayanna Howard: Human-Robot Interaction and Ethics of Safety-Critical Systems. Artificial Intelligence Podcast with Lex Fridman.
<https://www.youtube.com/watch?v=J21-7AsUcgM> More about Dr. Ayanna Howard: <https://howard.ece.gatech.edu/>

Mitigations to bias

Reviewing data

- Who created/organized/curated it?
- Are respected experts in the industry involved?
- What bias does it already have?

Consider system training

- Engage a diverse team, mindful of bias.
- What approach is being used?

Activating curiosity

Incorporating user experience (UX) research and human-computer interaction (HCI) methods to activate curiosity might involve

- engaging team members to take Implicit Association Test (Harvard University)
- trying Flip It to Test It activities for scenarios (HBR)
- performing Abusability Testing of system (Brown)
- creating “Black Mirror” Episodes (Fiesler)
(inspired by British dystopian sci-fi tv series of same name)

Potentially severe abuse and consequences exist, particularly for people who are frequently marginalized (women, racial and ethnic minorities, LGBTQ, etc.).

More methods to “Outsmart Your Own Biases.”: <https://hbr.org/2015/05/outsmart-your-own-biases>
Implicit Association Test (IAT): <https://implicit.harvard.edu/implicit/takeatest.html>

Humans ultimately remain in control.



*“Ensure humans can unplug
the machines”*

– Grady Booch, Scientist, Philosopher, IBM'er

TED Talk, Grady Booch, Scientist, Philosopher, IBM'er. https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Grand Challenge – Partner with the SEI

Catalyze the community in implementing ethics in AI engineering.

- Develop activities to activate curiosity, speculation, creativity, and imagination with regard to reducing ethical risks of systems (potential unintended/unwanted/inevitable consequences).
 - Create education and tools to support new and difficult work.
 - Experiment to prove what is most effective.
-
- Contact the SEI: Info@sei.cmu.edu