RSAConference2020
San Francisco | February 24 – 28 | Moscone Center

HUMAN ELEMENT

SESSION ID: MLAI-F03

Designing Trustworthy AI: A User Experience (UX) Framework

Carol J. Smith

Sr. Research Scientist - Human-Machine Interaction
Carnegie Mellon University's Software Engineering Institute
Twitter: @carologic @sei_etc

Legal

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.

DM20-0086

AI's Great Promise

- Empower us with knowledge
- Augment our effectiveness

- We can—and must—ensure that we keep humans safe and in control

RSAConference2020

Lack of Trust in Artificial Intelligence (AI)

Inadequate and abusive work

Assuming That Math Reduces Bias

Vetting applicant resumes

REUTERS

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 10 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Embedding Existing Bad Behaviors

Determining interest rates for mortgage lenders

Berkeley News

Research People Campus & community

BUSINESS & ECONOMICS, RESEARCH

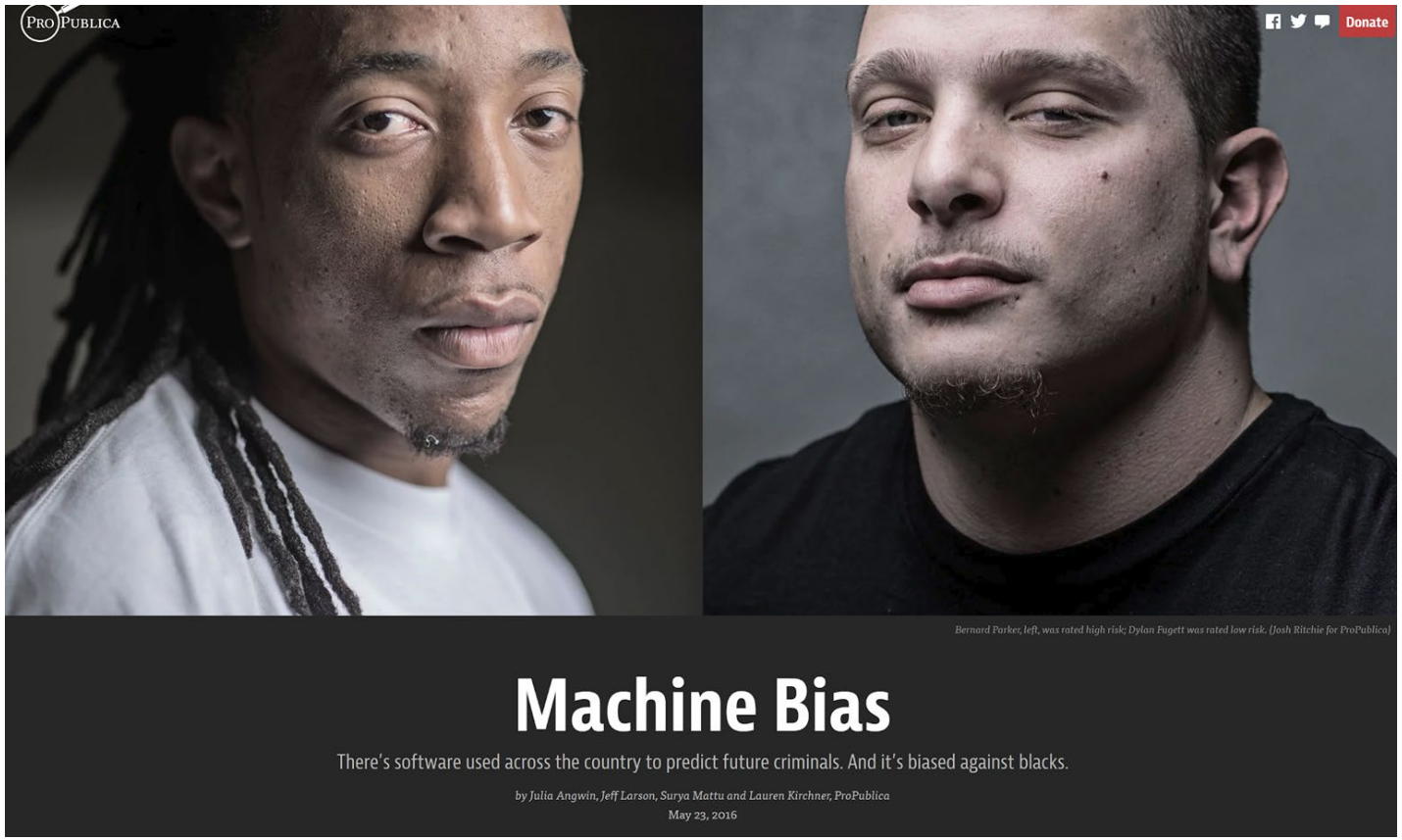
Mortgage algorithms perpetuate racial bias in lending, study finds

By Public Affairs, UC Berkeley | NOVEMBER 13, 2018

Tweet Share 150 Email Print

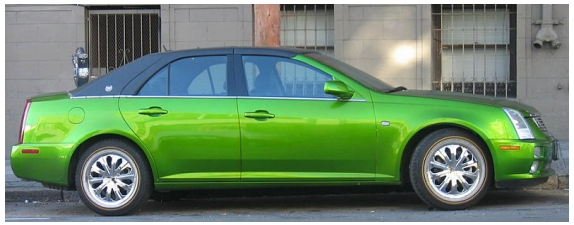
Dehumanizing thru Math

Predicting future criminal activity



Bias in Training Set

Training Set


Data Encountered




How can development teams create AI systems that are safe and trustworthy?

Today – Make Trustworthy AI Systems

- Technology ethics and bias
- Team qualities that lead to success
- User Experience (UX) Framework
- Checklist and Agreement to support work

RSAConference2020

Technology Ethics

Ethics

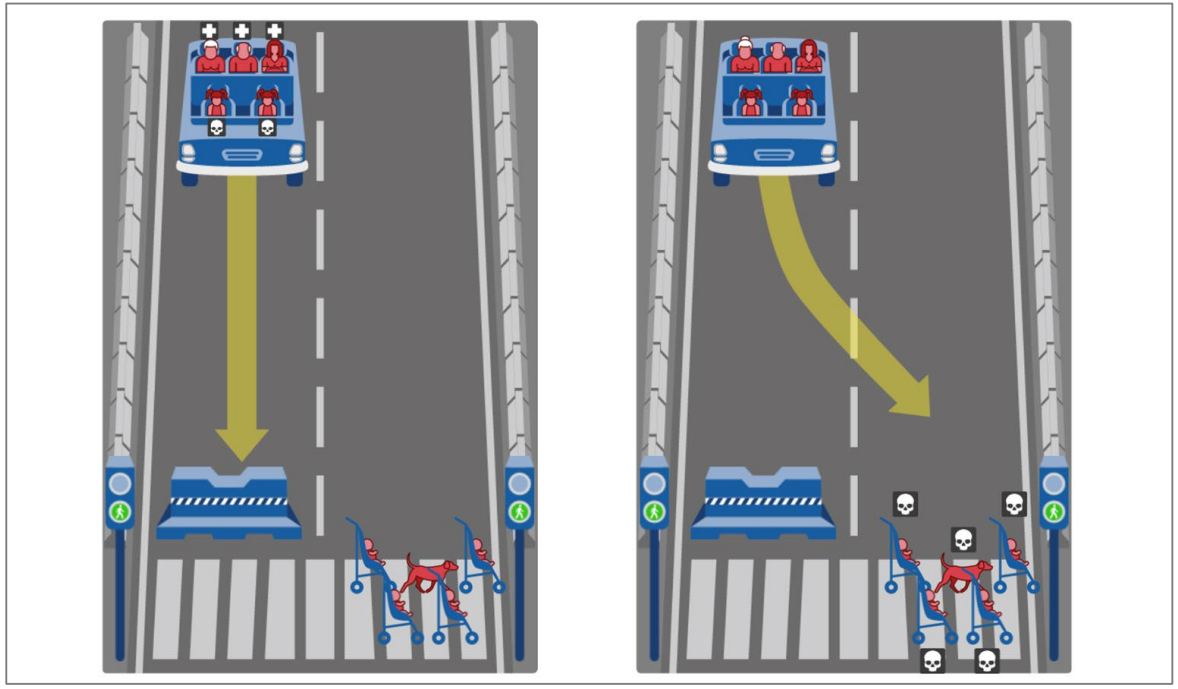
- Based on well-founded standards of right and wrong
- Standard of expected behavior that guides the correct course of action
- What impact does my work have?

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

Not Trolley Problems (hypotheticals in-the-moment)

MORAL MACHINE

MIT Moral Machine Project



Does the Trolley Problem Have a Problem?

What if your answer to an absurd hypothetical question had no bearing on how you behaved in real life?

By DANIEL ENGBER

JUNE 18, 2018 • 5:56 AM

Lisa Larson-Walker

Images: 1) MIT Moral Machine Project: http://moralmachine.mit.edu/
2) Does the Trolley Problem Have a Problem? https://slate.com/technology/2018/06/psychologys-trolley-problem-might-have-a-problem.html

This is early, purposeful work

- Not just about code
- Challenging work about conduct, interactions, intents
- Start with technology ethics
- Benefits
  - Harmonize cultural variations
  - Balance to pace of change, industry pressure
  - Explicit permission to consider and question breadth of implications

Adopt Technology Ethics

- What do you value?
- What lines won't you cross?
- Track progress

Association for Computing Machinery

AINOW INSTITUTE

Microsoft

Google

< >
Montréal Declaration
Responsible AI_
</ >
An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

ETHICAL OS

NETWORKS / CYBER, PENTAGON

Ethical AI For War? Defense Innovation Board Says It Can Be Done

The military figured out how to use nuclear power safely, the advisory panel said, and it can do the same with artificial intelligence.

By SYDNEY J. FREEDBERG JR. on October 31, 2019 at 12:21 PM

Textron Ripsaw M5 robot in an armed configuration

NETWORKS / CYBER, PENTAGON

Ethical AI For War? Defense Innovation Board Says It Can Be Done

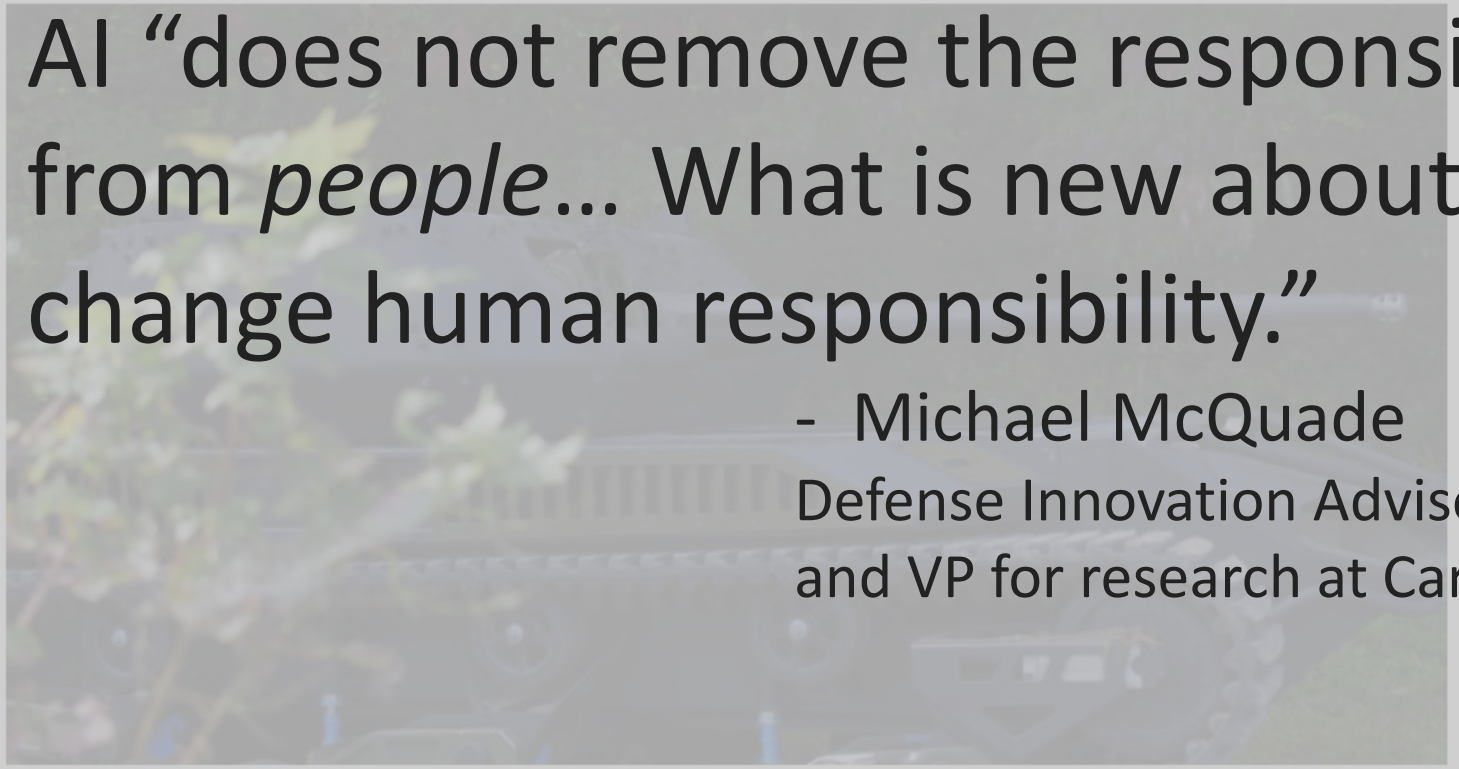
The military figured out how to use nuclear power safely, the advisory panel said, and it can do the same with artificial intelligence.

By SYDNEY J. FREEDBERG JR. on October 31, 2019 at 12:21 PM

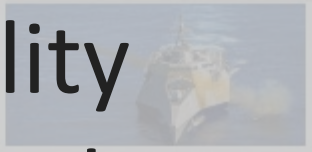
AI "does not remove the responsibility from *people*… What is new about AI does not change human responsibility."

- Michael McQuade

Defense Innovation Advisory Board member and VP for research at Carnegie Mellon University



Textron Ripsaw M5 robot in an armed configuration



NAVAL WARFARE, SPONSORED

When Innovation Succeeds

Raytheon's Naval Strike Missile will play an important role for the U.S. Navy. Learn how the USS Gabrielle Giffords became the first navy littoral combat ship to launch the NSM in an integrated setup.

Recommended

"When you know better, you do better."

- Betty Easterly

Be aware and informed.

RSAConference2020

Bias

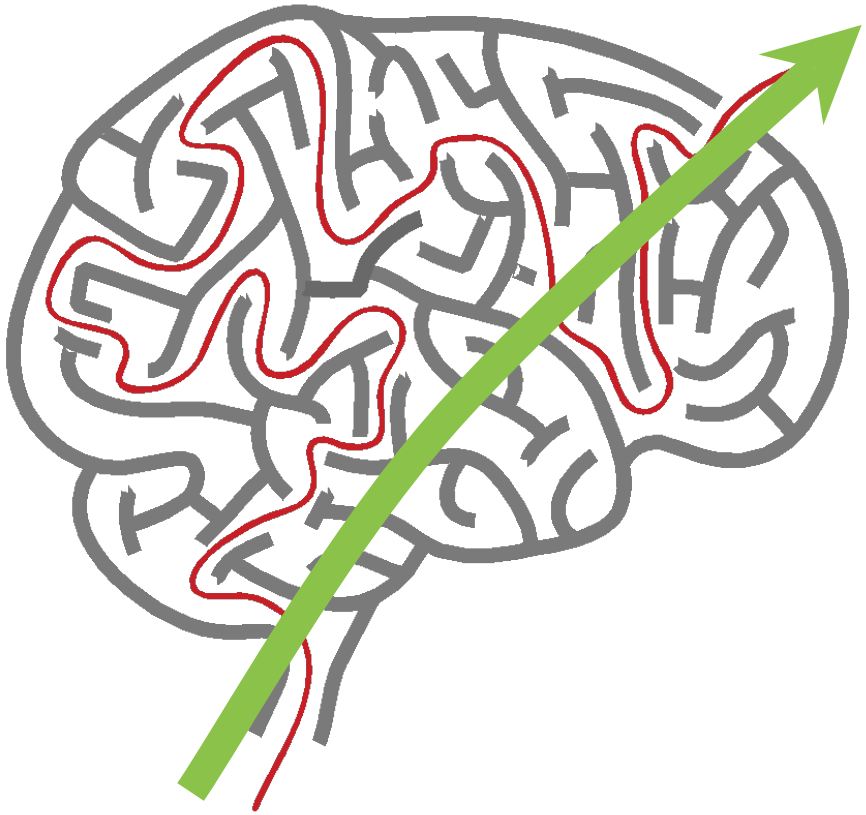
Data Design Solution

...are biased

Due to team member's…

- Social class
- Resource availability
- Education
- Race, gender, sexuality
- Culture, theology, tradition
- More…

Bias: Human's Mental Shortcut



- Not inherently bad, may be misapplied
- Implicit = invisible (intuition)
- Not necessarily in sync with conscious beliefs
- **Can be managed and changed**

Graphic by Karen Bachmann as presented in:
https://www.slideshare.net/karenbachmann/know-thyself-understanding-and-managing-biases

Start with building awareness and knowing ourselves

How are our biases affecting our work?

RSAConference2020

Team Qualities for Trustworthy AI

What is a diverse team?

Gender, race, culture,
education (school, program, etc.),
thinking process,
disability status, and more…

Diverse, Talented and Multi-Disciplinary

Includes skill set and problem framing approach

- Data Scientists
- Machine learning experts
- Programmers
- System architects
- Product managers, etc.

- Not lowering bar – extending it

(context dependent)

Curiosity Experts

- Understand
  - Situation
  - Abilities of people who will use system
  - How system will be used
- Activate curiosity in others via UX activities
- Social scientists, ethicists and more…

Why diverse teams?

- Focus more on facts
- Process facts more carefully
- More innovative

- "…become more aware of their own potential biases"

Harvard Business Review

Why Diverse Teams Are Smarter. Harvard Business Review.
https://hbr.org/2016/11/why-diverse-teams-are-smarter

Great Minds Think Different

Inclusive Workplace Requirements

- Representatively diverse leadership = retention
- Individuals differences are acknowledged and accepted
  - Bring "whole selves" to work
  - Welcoming community environment
  - Feel valued, connected, that we belong

Coalesce on Shared Set of Technology Ethics

- 1. Well-being
- 2. Respect for autonomy
- 3. Protection of privacy and intimacy
- 4. Solidarity
- 5. Democratic participation
- 6. Equity
- 7. Diversity inclusion
- 8. Prudence
- 9. Responsibility
- 10. Sustainable development

Montréal Declaration for a responsible development of artificial intelligence.
https://www.montrealdeclaration-responsibleai.com/the-declaration

Diverse, inclusive leaders

Diverse, Multi-Disciplinary Teams

Shared Tech Ethics

RSAConference2020

UX Framework

Designing Trustworthy AI

How do we get there?

Trustable,
Ethical AI

Montréal Declaration for a responsible development of artificial intelligence.
https://www.montrealdeclaration-responsibleai.com/the-declaration

Conversations for Understanding

- UX Framework guides AI development teams
- 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?

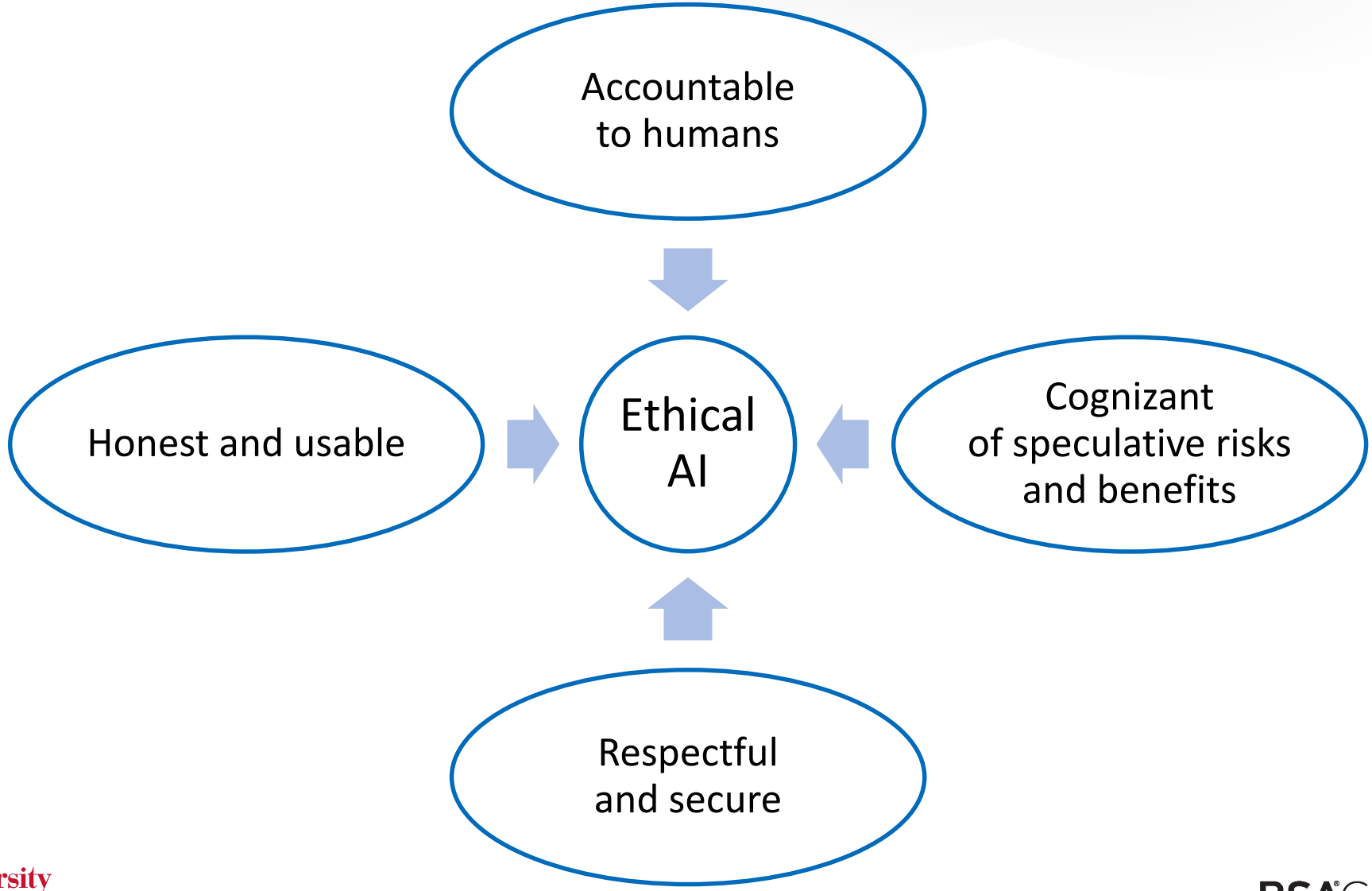
*"How is this ML model shifting power?" @riakall #NeurIPS2019

"Be uncomfortable"

- Laura Kalbag

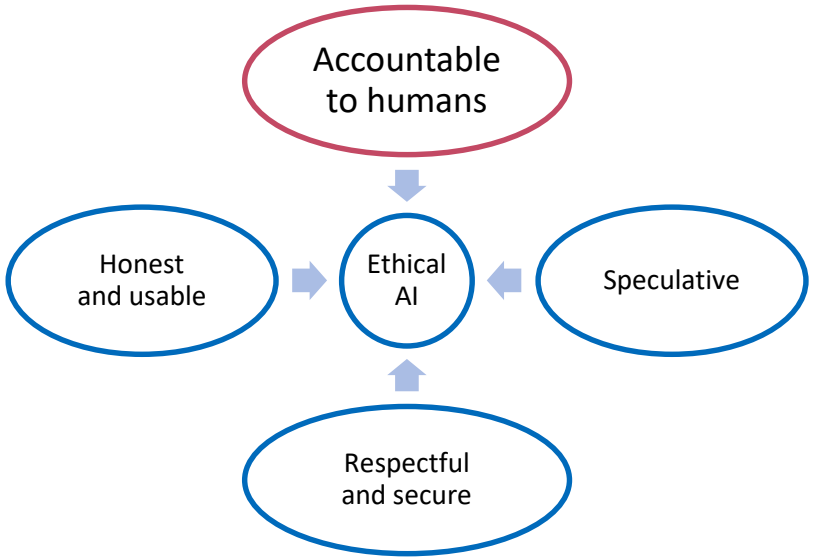
Ethical design is not superficial.

UX Framework for Designing Trustworthy AI



Accountable to Humans

- Ensure humans have ultimate control
  - Able to monitor and control risk
- Human responsibility for final decisions
  - Person's life
  - Quality of life
  - Health
  - Reputation



“Ensure humans can unplug the machines”

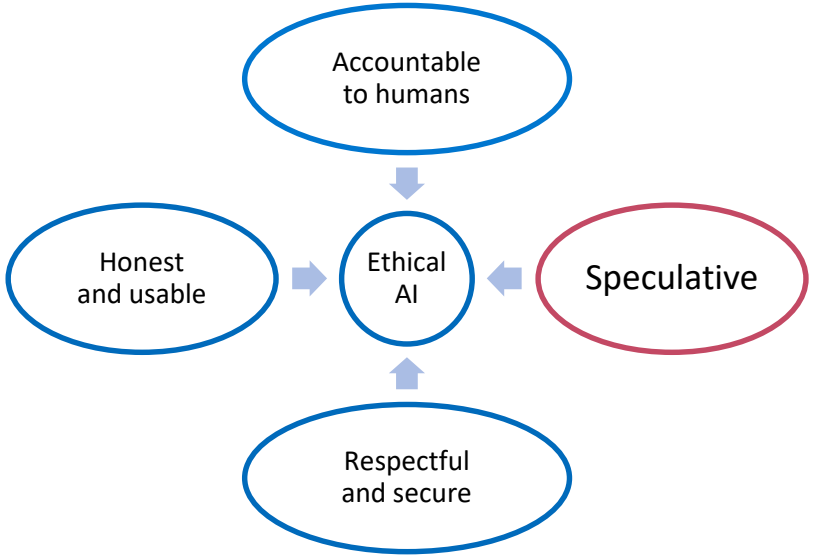
– Grady Booch

TED Talk, Grady Booch, Scientist, Philosopher, IBM’er

https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Cognizant of Speculative Risks and Benefits

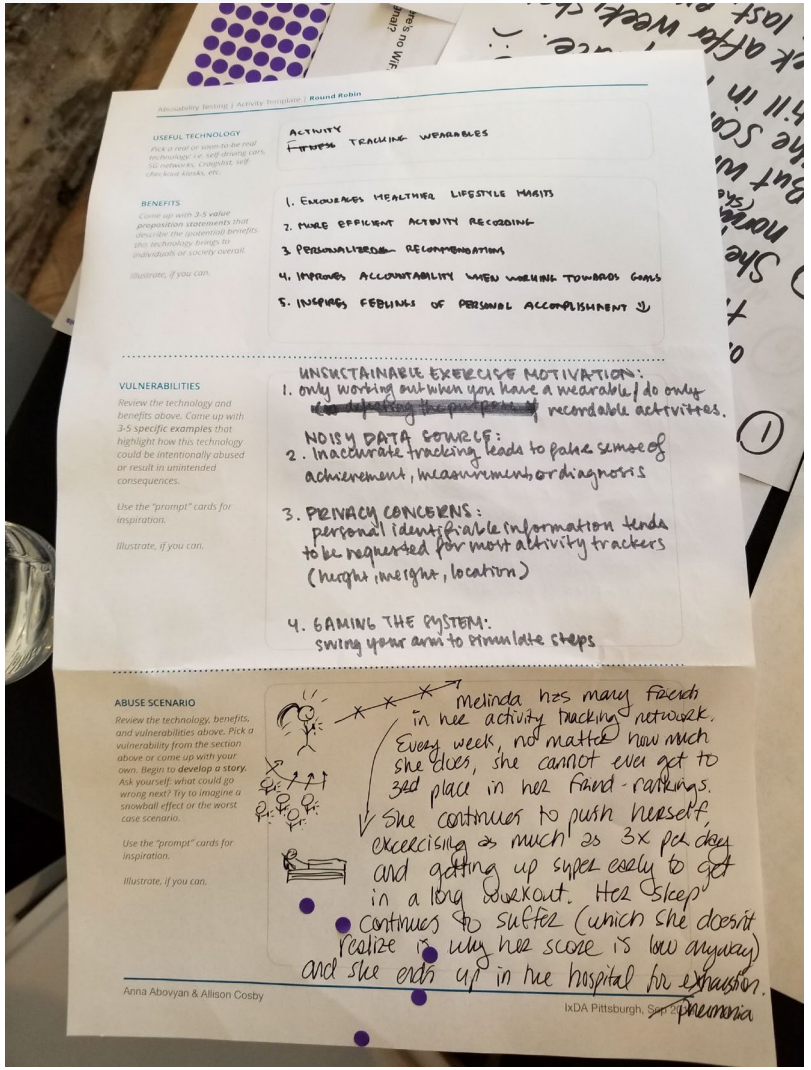
- Identify full range of
  - Harmful, malicious use, as well as good, beneficial use
  - Blind spots and unwanted/unintended consequences



Speculative: Conduct UX research and activate curiosity

Abusability Testing

- Speculate about misuse and abuse
- Create “Black Mirror” episodes
  - Severe abuse and consequences

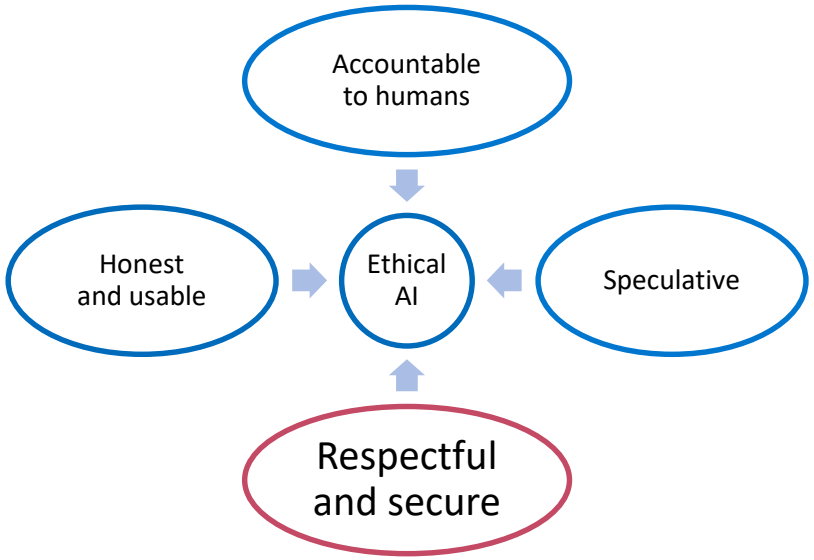


Speculative: Create communication & mitigation plans

- Plan for unwanted consequences
- Misuse and abuse of AI system
  - Who can report?
  - To whom?
  - Turn off?
  - Who notified?
  - Consequences?

Respectful and Secure

- Values of humanity, ethics, equity, fairness, accessibility, diversity and inclusion
- Respect privacy and data rights
- Make system robust, valid and reliable
- Provide understandable security

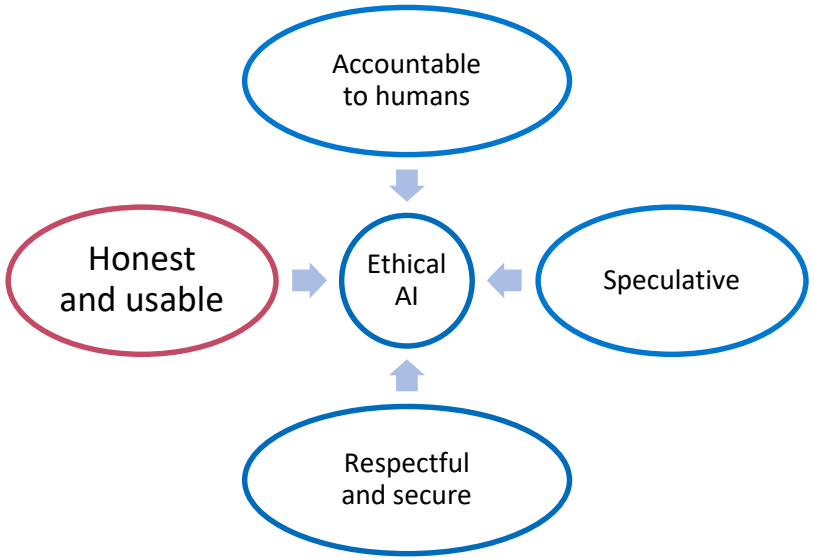


Fair: Remove unwanted bias in data

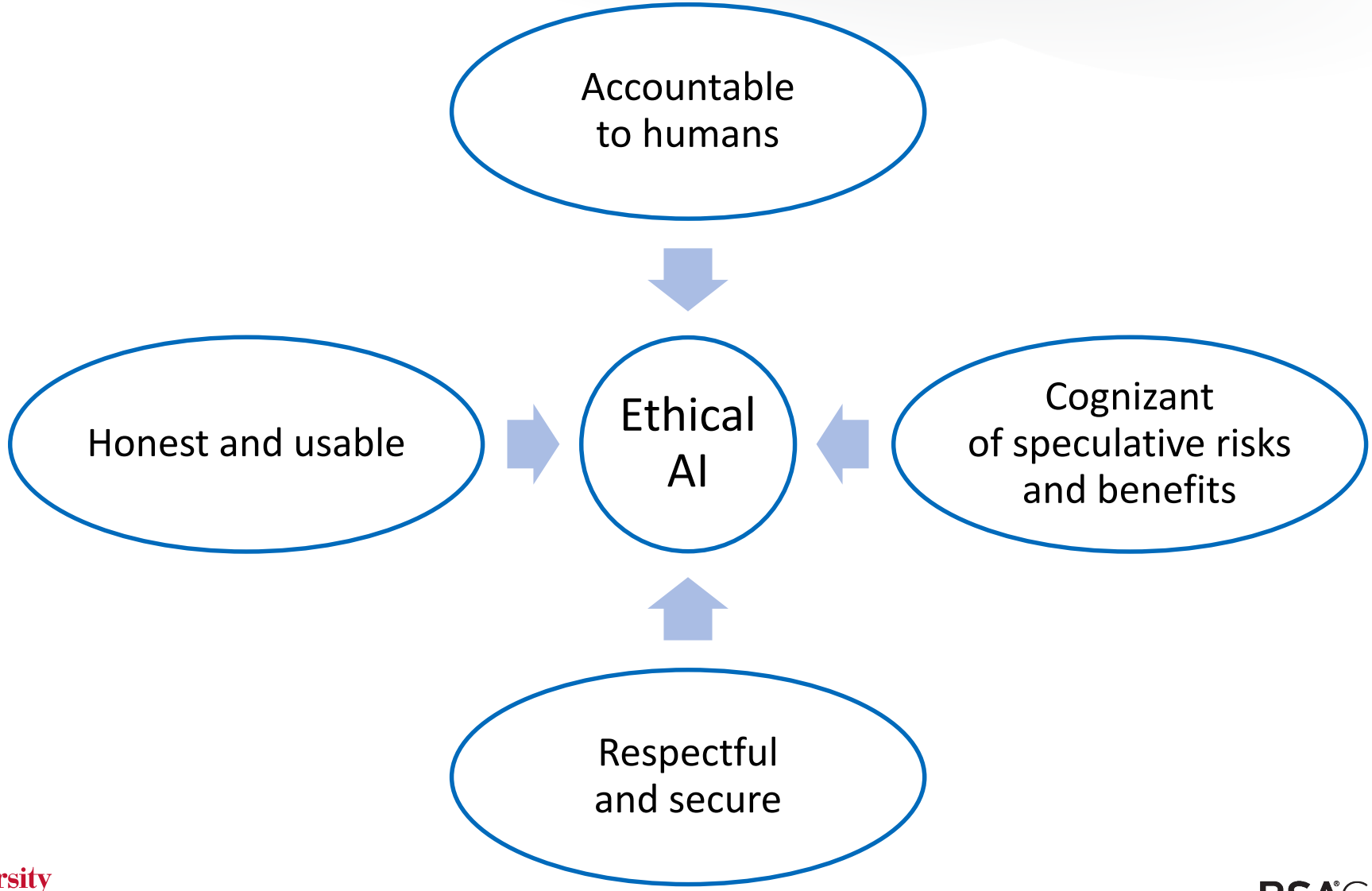
- Show awareness of known and desirable bias
- Acknowledge issues
- Overcommunicate on issues

Honest and Usable

- Value transparency with the goal of engendering trust
- Explicitly state identity as an AI system



UX Framework: Being intentional, keeping people safe



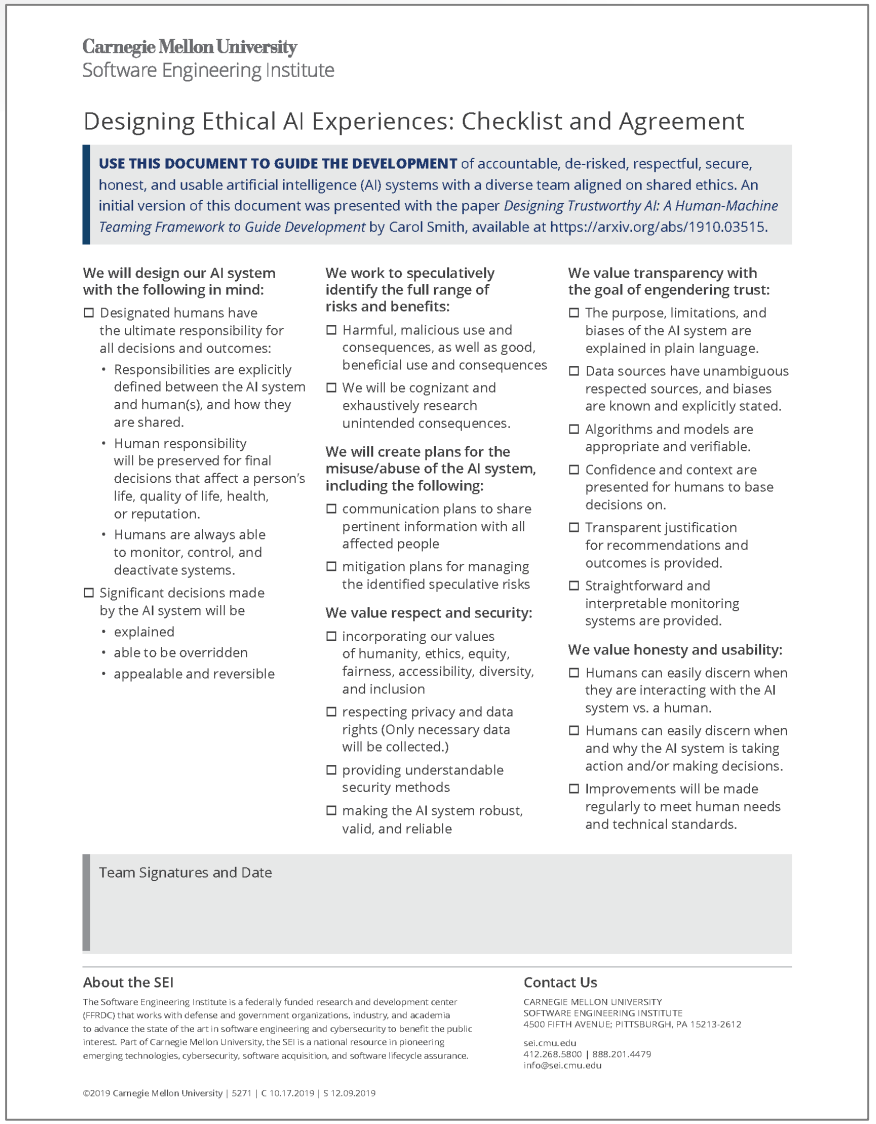
RSAConference2020

Checklist and Agreement

Designing Trustworthy AI

Checklist and Agreement

- Pair with Tech Ethics
  - Bridge gap between "do no harm" and reality
- Reduce risk and unwanted bias
- Mitigation planning
- Support inspection



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

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

Prompts for Conversations

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.

Team Signatures and Date

Reward team members for finding ethics bugs

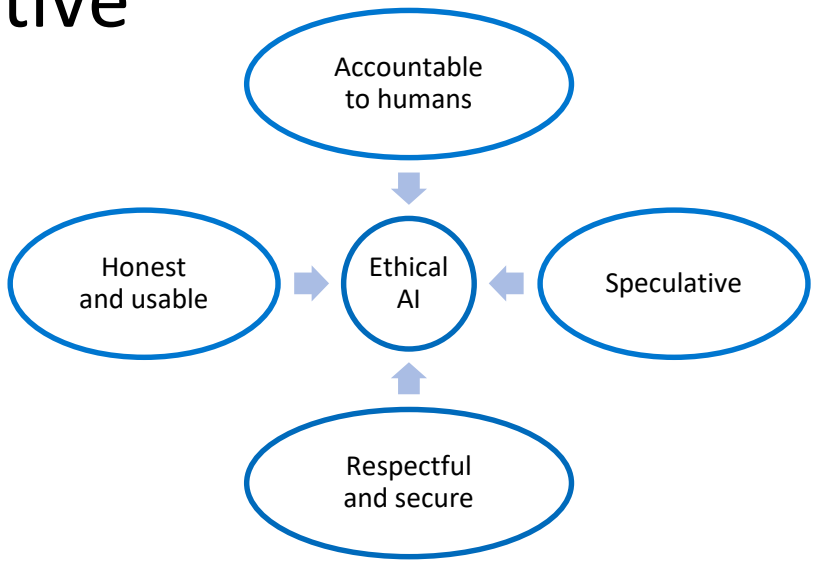
- Dr. Ayanna Howard
  - on the Artificial Intelligence Podcast with Lex Fridman

Evangelize for human values

Ethical. Transparent. Fair.

We aren't perfect, AI won't be perfect

- Diverse teams, inclusive environments
- Adopt tech ethics
- Encourage deep conversations (Checklist)
- Activate curiosity; be speculative; imaginative
- Make trustworthy AI systems



Apply What You Have Learned Today

- Next week:
  - Identify initial set of technology ethics
  - Introduce tech ethics and Checklist to 1 team*
- 3 months:
  - Formally adopt technology ethics based on learnings
  - Run 1+ abusability workshop
  - Expand use of Checklist
- 6 months:
  - Measure progress (e.g. proactiveness, quality conversations vs. surprises)

*Experiment to learn

RSAConference2020

San Francisco | February 24 – 28 | Moscone Center

HUMAN ELEMENT

SESSION ID: **MLAI-F03**

Designing Trustworthy AI: A UX Framework

Carol J. Smith, cjsmith@sei.cmu.edu

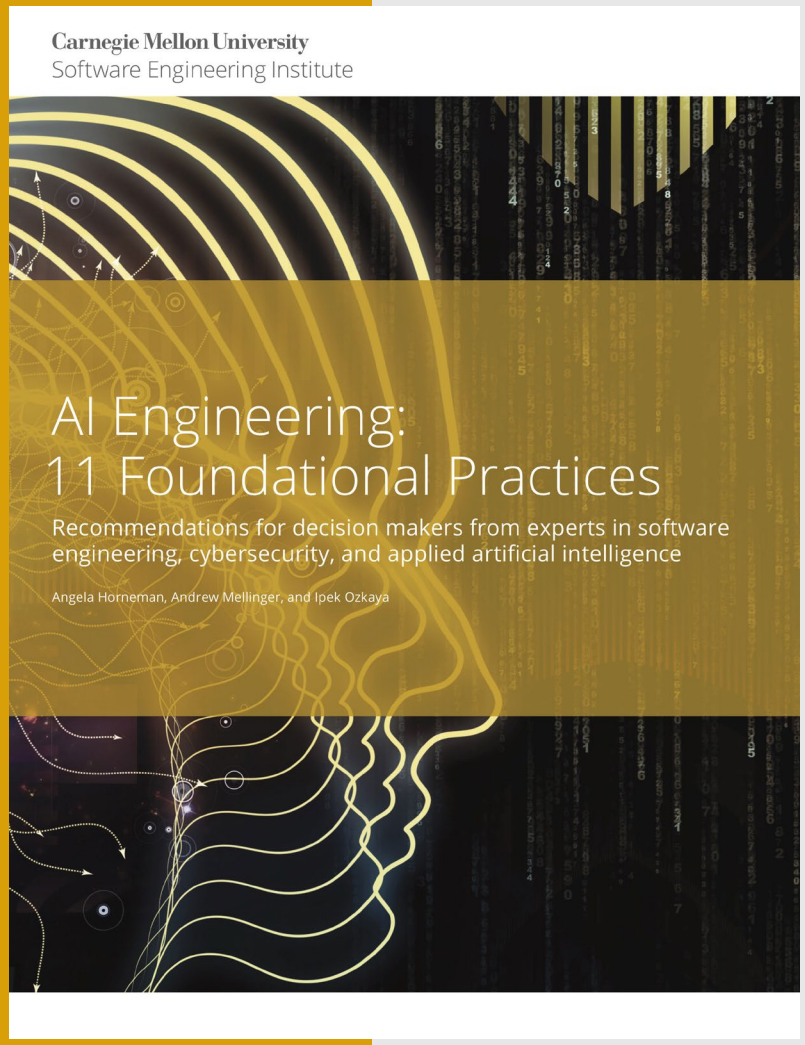
Sr. Research Scientist - Human-Machine Interaction
Carnegie Mellon University's Software Engineering Institute
Twitter: @carologic @sei_etc

SEI on HMT: https://sei.cmu.edu/research-capabilities/all-work/display.cfm?customel_datapageid_4050=197910

RSAConference2020

Resources

Designing Trustworthy AI



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

resources.sei.cmu.edu/library/asset-view.cfm?assetid=633647

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

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

Resources

- Black Mirror, Light Mirror: Teaching Technology Ethics Through Speculation. Casey Fiesler. Oct 15, 2018. https://howwegettonext.com/the-black-mirror-writers-room-teaching-technology-ethics-through-speculation-f1a9e2deccf4
- UX in the Age of Abusability. The role of Composition, Collaboration, and Craft in building ethical products. Dan Brown. Sep 18, 2018. https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13
- AI Now 2017 Report, New York University, and AI Now https://assets.ctfassets.net/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf
- "How IBM is Competing with Google in AI." The Information. https://www.theinformation.com/how-ibm-is-competing-with-google-in-ai?eu=2zIDMNYNjDp7KqL4YqAXXA
- "The business case for augmented intelligence" by Nancy Pearson, VP Marketing, IBM Cognitive. https://medium.com/cognitivebusiness/the-business-case-for-augmented-intelligence-36afa64cd675#.qqzvunakw
- "Inside Intel: The Race for Faster Machine Learning" http://www.intel.com/content/www/us/en/analytics/machine-learning/the-race-for-faster-machine-learning.html

More Resources

- “Update: Why this week’s man-versus-machine Go match doesn’t matter (and what does)” by Dana Mackenzie. Science Magazine. Mar. 15, 2016 http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does
- “For IBM’s CTO for Watson, not a lot of value in replicating the human mind in a computer.” by Frederic Lardinois (@fredericl), TechCrunch, Posted Feb 27, 2017. https://techcrunch.com/2017/02/27/for-ibms-cto-for-watson-not-a-lot-of-value-in-replicating-the-human-mind-in-a-computer/
- “Google and IBM: We Want Artificial Intelligence to Help You, Not Replace You” Most Powerful Women by Michelle Toh. Mar 02, 2017. Fortune. http://fortune.com/2017/03/02/google-ibm-artificial-intelligence/
- “Facebook scales back AI flagship after chatbots hit 70% f-AI-lure rate - 'The limitations of automation‘” by Andrew Orlowski. Feb 22, 2017. The Register https://www.theregister.co.uk/2017/02/22/facebook_ai_fail/
- “Microsoft is deleting its AI chatbot's incredibly racist tweets” by Rob Price. Mar. 24, 2016. Business Insider UK. http://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3

Even More Resources

- “IBM’s Automated Radiologist Can Read Images and Medical Records” by Tom Simonite, February 4, 2016. Intelligent Machines, MIT Technology Review. https://www.technologyreview.com/s/600706/ibms-automated-radiologist-can-read-images-and-medical-records/
- “The IBM, Salesforce AI Mash-Up Could Be a Stroke of Genius” by Adam Lashinsky, Mar 07, 2017. Fortune. http://fortune.com/2017/03/07/data-sheet-ibm-salesforce/
- "Google can now tell you're not a robot with just one click" by Andy Greenberg. Dec. 3, 2014. Security: Wired. https://www.wired.com/2014/12/google-one-click-recaptcha/
- “Essentials of Machine Learning Algorithms (with Python and R Codes)” by Sunil Ray, August 10, 2015. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/
- IBM on Machine Learning https://www.ibm.com/analytics/us/en/technology/machine-learning/
- “At Davos, IBM CEO Ginni Rometty Downplays Fears of a Robot Takeover” by Claire Zillman, Jan 18, 2017. Fortune. http://fortune.com/2017/01/18/ibm-ceo-ginni-rometty-ai-davos/
- “Google and IBM: We Want Artificial Intelligence to Help You, Not Replace You” by Michelle Toh. Mar 02, 2017. Fortune. http://fortune.com/2017/03/02/google-ibm-artificial-intelligence/

Yes, even more resources

- Video: “IBM Watson Knowledge Studio: Teach Watson about your unstructured data” https://www.youtube.com/watch?v=caIdJjtvX1s&t=6s
- “The optimist’s guide to the robot apocalypse” by Sarah Kessler, @sarahfkessler. March 09, 2017. QZ. https://qz.com/904285/the-optimists-guide-to-the-robot-apocalypse/
- “AI Influencers 2017: Top 30 people in AI you should follow on Twitter" by Trips Reddy @tripsy, Senior Content Manager, IBM Watson . February 10, 2017 https://www.ibm.com/blogs/watson/2017/02/ai-influencers-2017-top-25-people-ai-follow-twitter/
- “3 guiding principles for ethical AI, from IBM CEO Ginni Rometty” by Alison DeNisco. January 17, 2017, Tech Republic http://www.techrepublic.com/article/3-guiding-principles-for-ethical-ai-from-ibm-ceo-ginni-rometty/
- "Transparency and Trust in the Cognitive Era" January 17, 2017 Written by: IBM THINK Blog https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/
- "Ethics and Artificial Intelligence: The Moral Compass of a Machine“ by Kris Hammond, April 13, 2016. Recode. http://www.recode.net/2016/4/13/11644890/ethics-and-artificial-intelligence-the-moral-compass-of-a-machine

Last bit: I promise

- "The importance of human innovation in A.I. ethics" by John C. Havens. Oct. 03, 2015 http://mashable.com/2015/10/03/ethics-artificial-intelligence/#yljsShvAFsqy
- "Me, Myself and AI" Fjordnet Limited 2017 - Accenture Digital. https://trends.fjordnet.com/trends/me-myself-ai
- "Testing AI concepts in user research" By Chris Butler, Mar 2, 2017. https://uxdesign.cc/testing-ai-concepts-in-user-research-b742a9a92e55#.58jtc7nzo
- "CMU prof says computers that can 'see' soon will permeate our lives“ by Aaron Aupperlee. March 16, 2017. http://triblive.com/news/adminpage/12080408-74/cmu-prof-says-computers-that-can-see-soon-will-permeate-our-lives