Carnegie Mellon University
Software Engineering Institute

Academic Research

Trustworthy AI Systems

TALARIA SUMMER INSTITUTE, JULY 2023

Carol J. Smith
Carnegie Mellon University, Software Engineering Institute
Sr. Research Scientist, Human-Machine Interaction

Copyright Statement

Carol J. Smith

Carnegie Mellon University Software Engineering Institute

Software Engineering Institute

Human-Computer Interaction Institute

AI Division Staff

- Sr. Research Scientist, human-machine interaction
- AI/ML, autonomy, emerging technologies
- Government agencies

Adjunct Instructor

- Interaction Design Overview
- Human-centered design
- Prototyping
- Design and iteration

Research to reduce risk and avoid waste

Carnegie Mellon University Software Engineering Institute

Software and AI systems

- Decision support
- Recommender systems
- Large Language Models (LLMs)
- Voice recognition (smart speakers)
- Speech to text/text to speech (live captions)
- And more…

Carnegie Mellon University Software Engineering Institute

"The biggest waste of all is building something no one wants"

Eric Ries @ericries via @MelBugai on Twitter at LeanStartupMI in 2011

Everything is Designed

Carnegie Mellon University
Software Engineering Institute

Reduce Risk

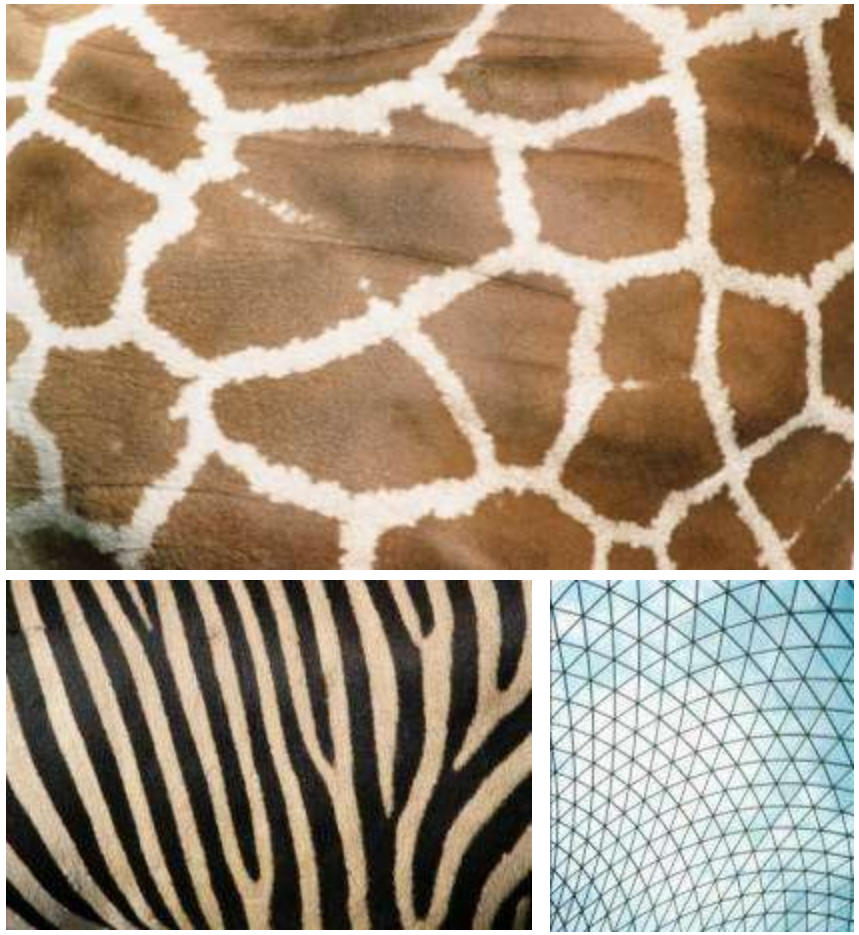
https://www.reddit.com/r/dangerousdesign/comments/3s75gz/cooking_spray_insect_killer/

Consider changes over time

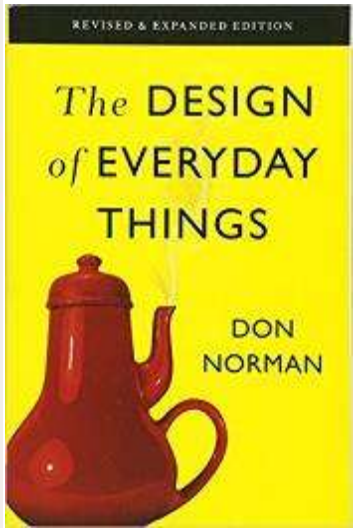
https://boingboing.net/2014/02/25/the-story-behind-the-too-coo.html

Long tradition – applying helpful patterns

- Human-Computer Interaction
- User Experience
- Usability
- Interaction Design

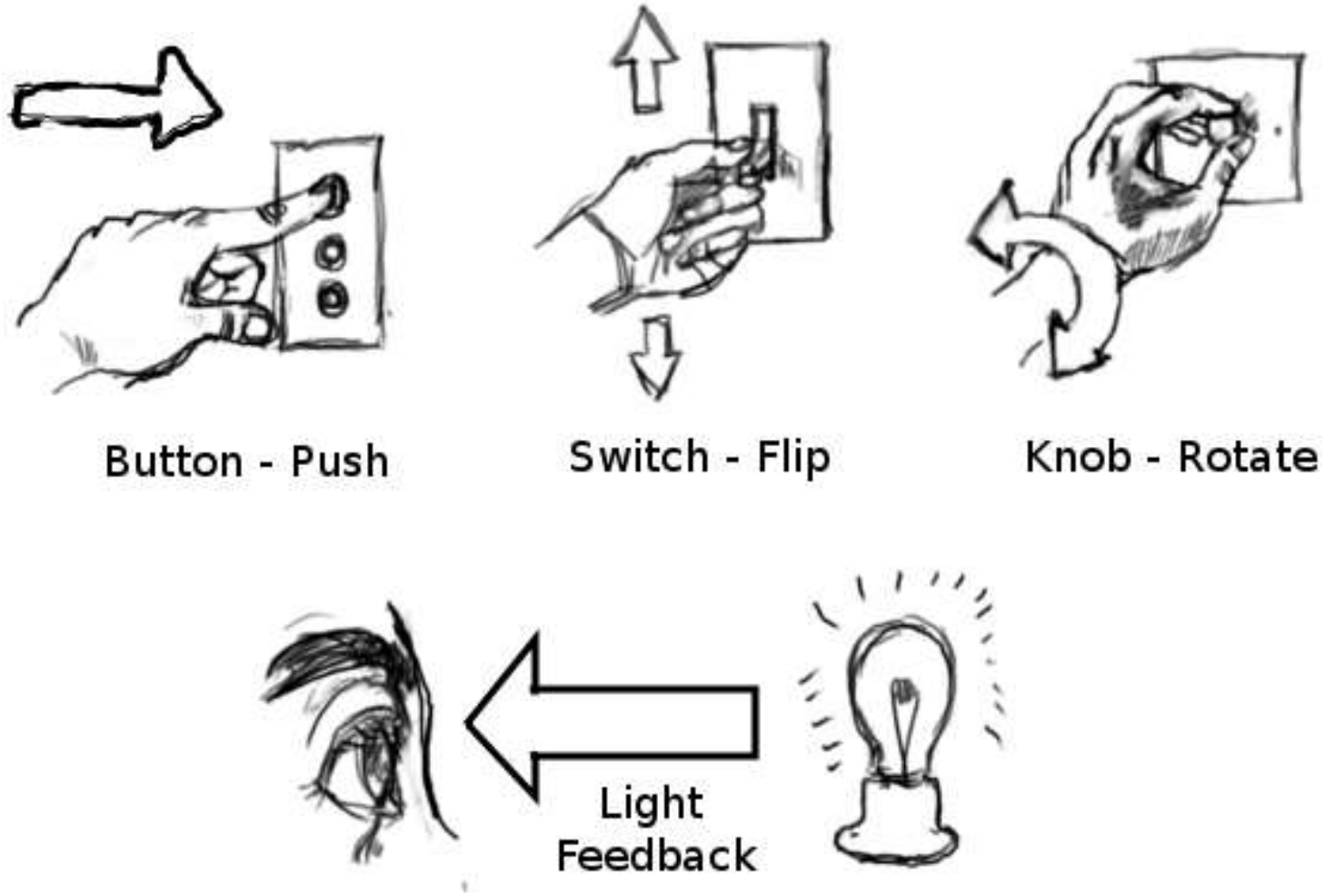


Design Better Experiences



The "masochist's teapot", as borrowed by Don Norman for his book *The Design of Everyday Things*.
Image: http://playerside.blogspot.com/2012/12/the-design-of-everyday-gaming-feedback.html

Make informed design choices



Drawings of Affordance: http://paaralan.blogspot.com/2010/09/affordance-and-educational-games.html

My Work

Carnegie Mellon University Software Engineering Institute

Focus on interesting and influential topics

- Someone has an idea
- Explore options
- Plan a research project

Projects are long (1+ years)

Work with software engineers, machine learning researchers, scientists, and more…

Explore the situation

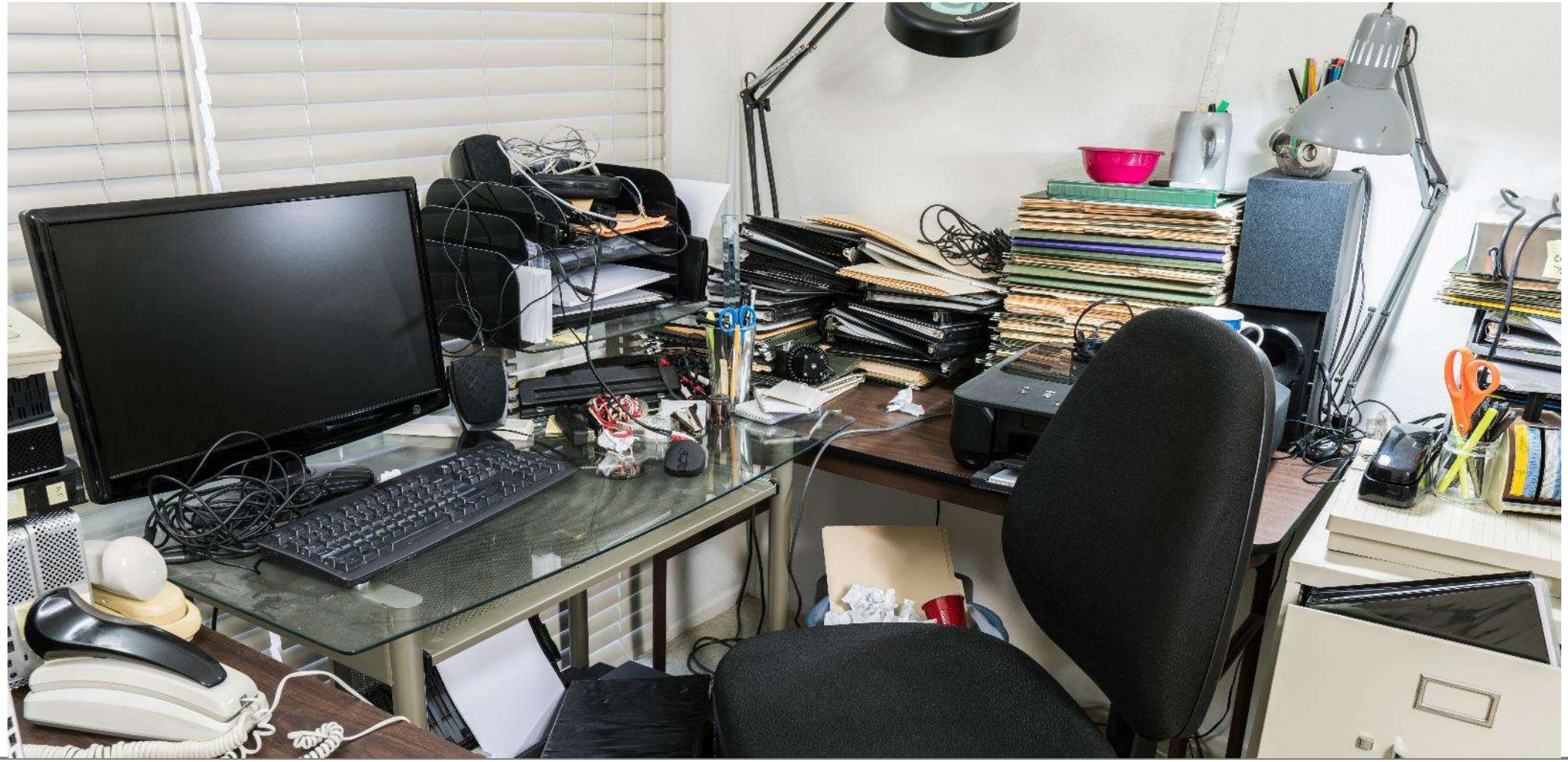
- Who would use the system?
- What would they do?
- What are their needs?

- Conduct Research Observations, Interviews

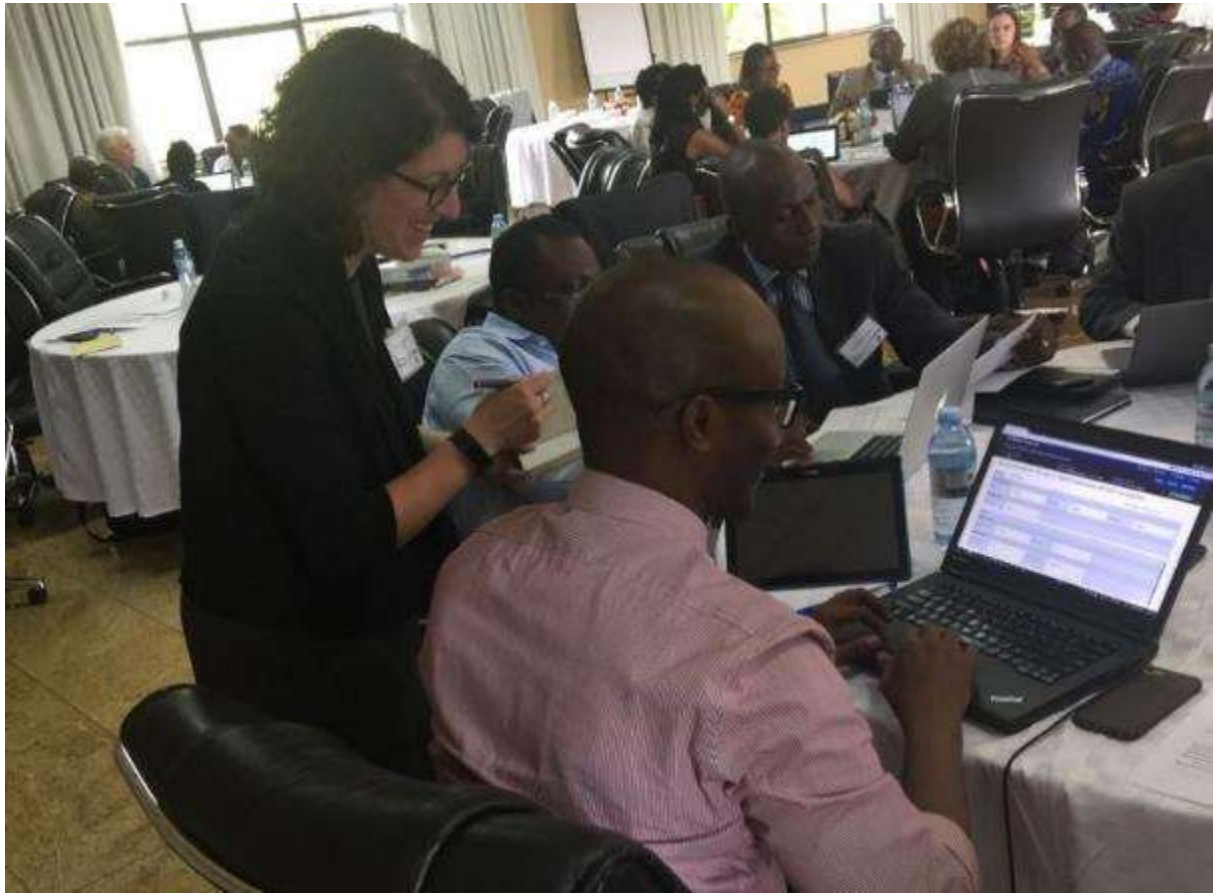
Observations…

- Learn real processes - existing product's actual use
- Workarounds
- Number and type of interruptions
- Remote considerations
- Understand how processes fit together
- What information is lost between systems (print and retype)
- Identify key opportunities

Artifacts extend understanding



Observe use of system and environment

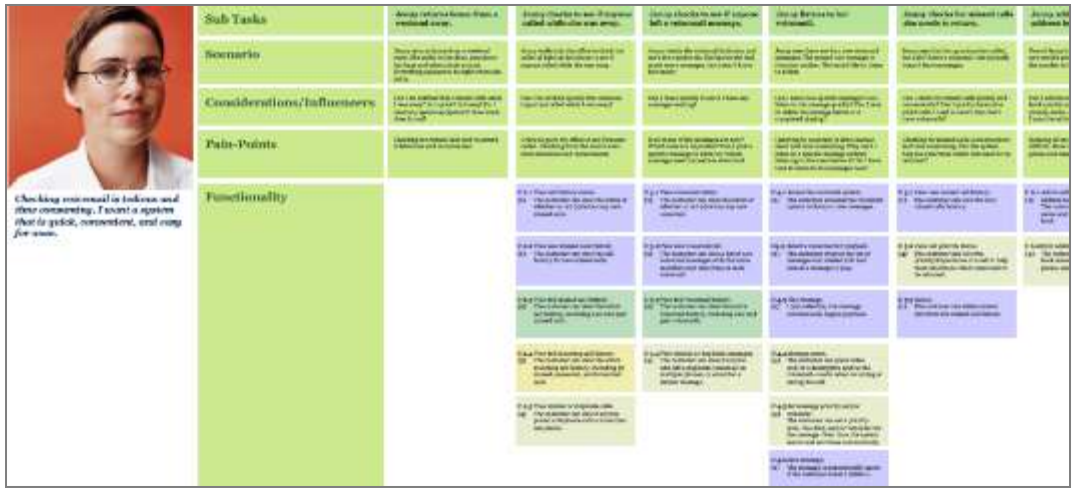


Gear

Analyze Information

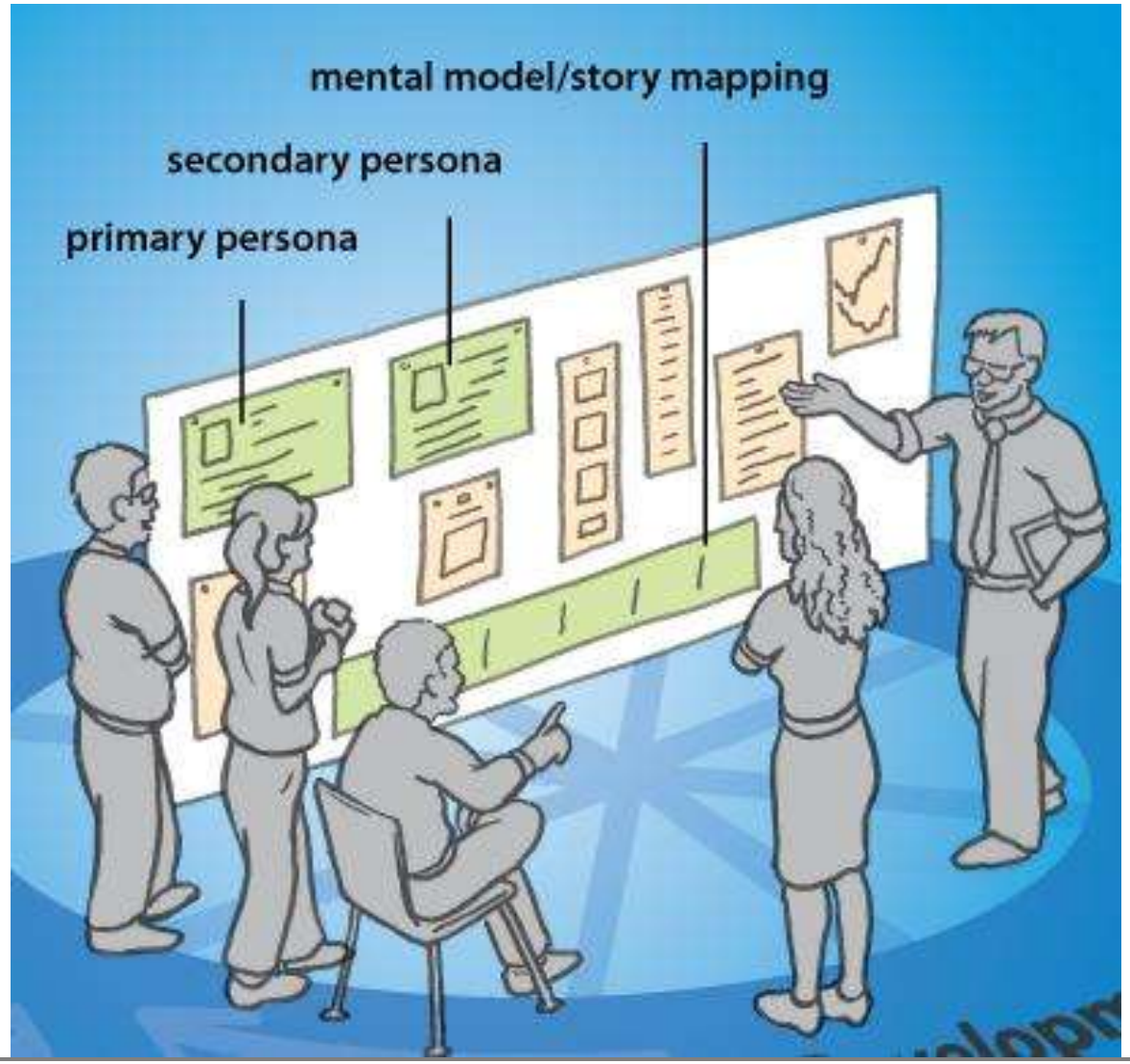
Map out information
- Look for patterns
- What people do
- Why they do it?

- Create helpful references
- Not “books” for shelving



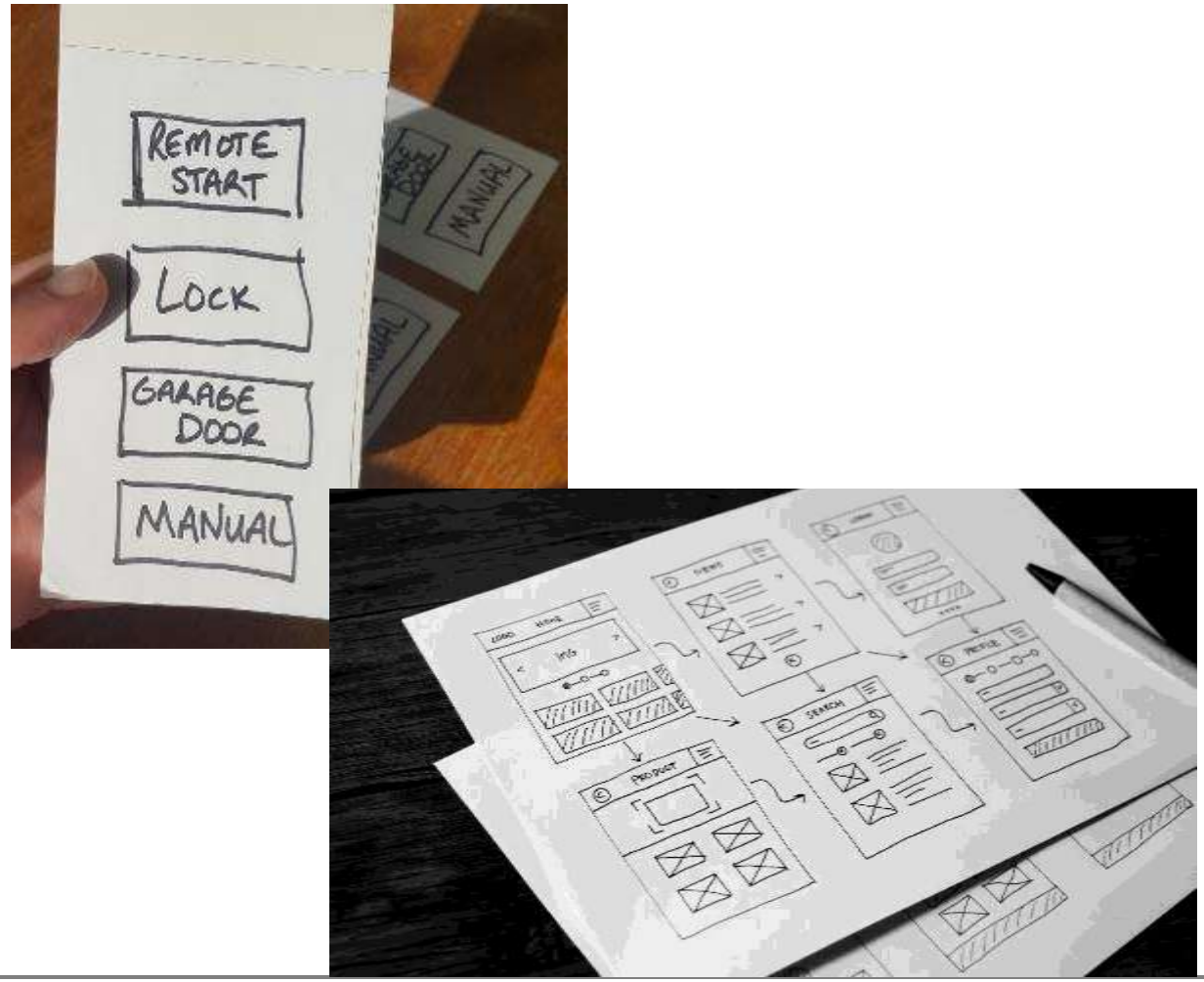
Example of a Task Analysis by Todd Zaki Warfel from his Agile2010 presentation "Opening the Kimono a look behind the design process."

Shared understanding and thoughtful discussion



Prototype – Experiments to Learn

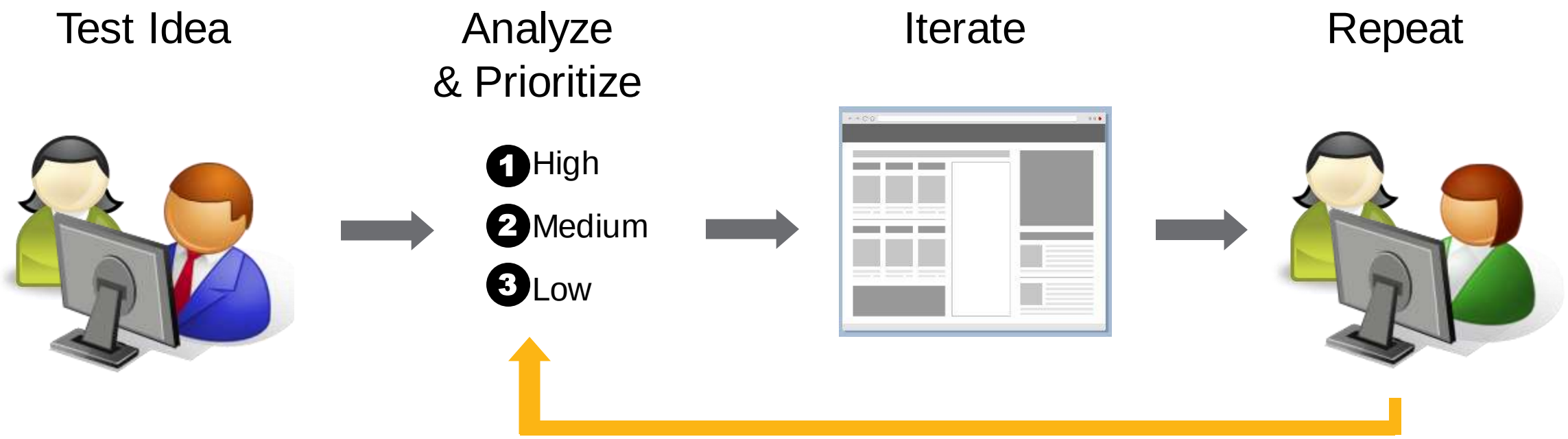
- How can I best support the user?
- How do I solve this problem?
- What interaction will meet the end users' needs?



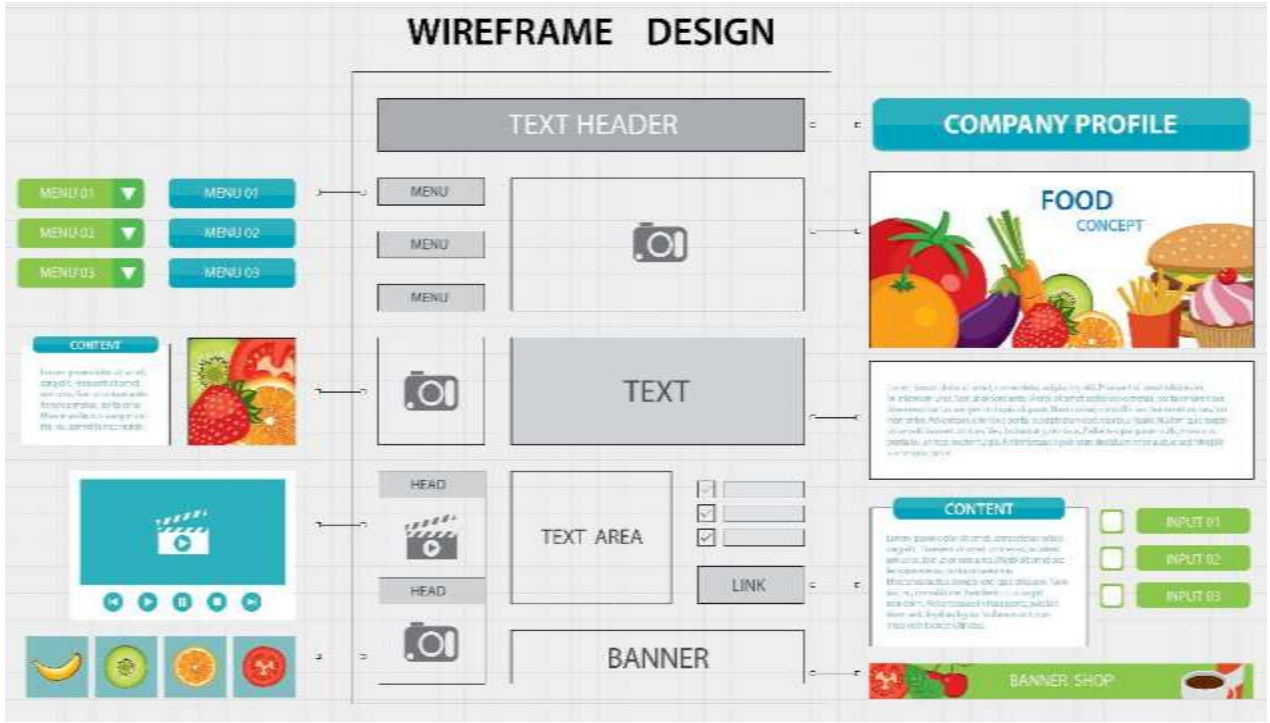
Early Learning: Observe use of prototype

Image from Steve Krug's "Rocket Surgery Made Easy"

Iterative Cycles: Feedback and Improvement



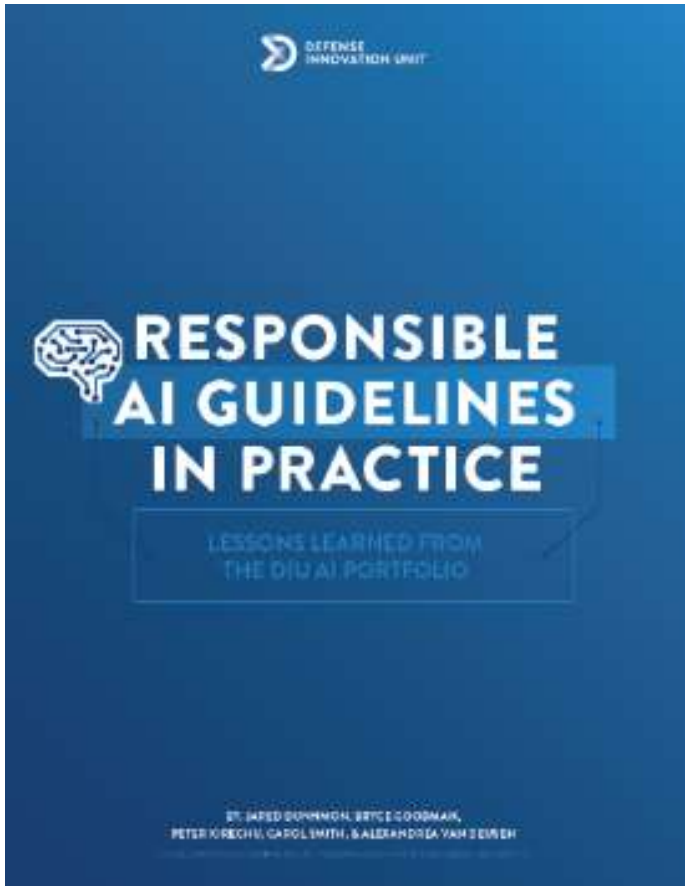
Work with Visual Designers

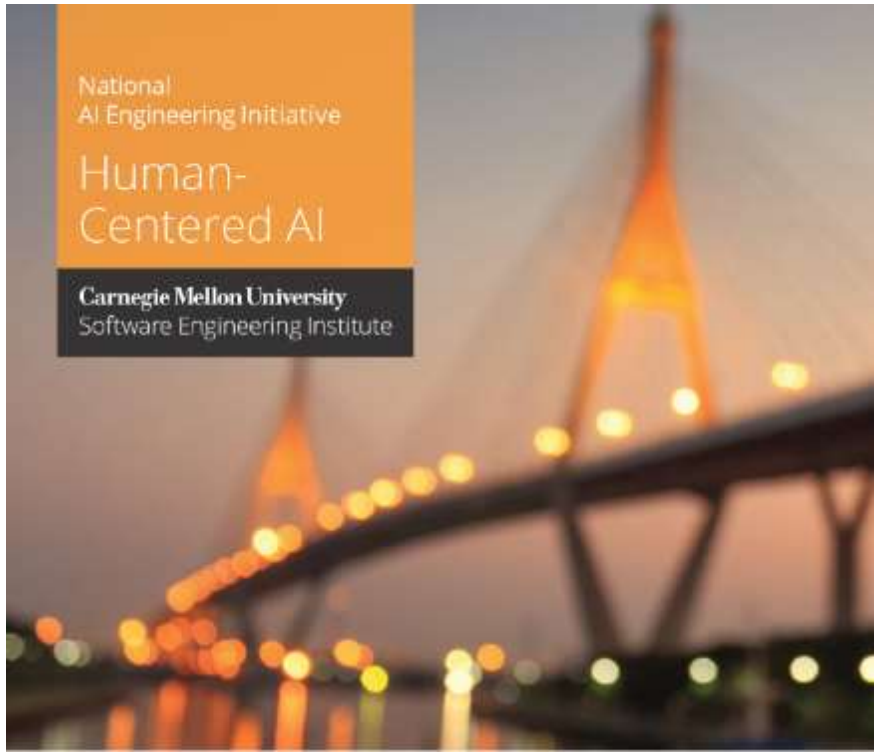


Presentations and Publications


Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

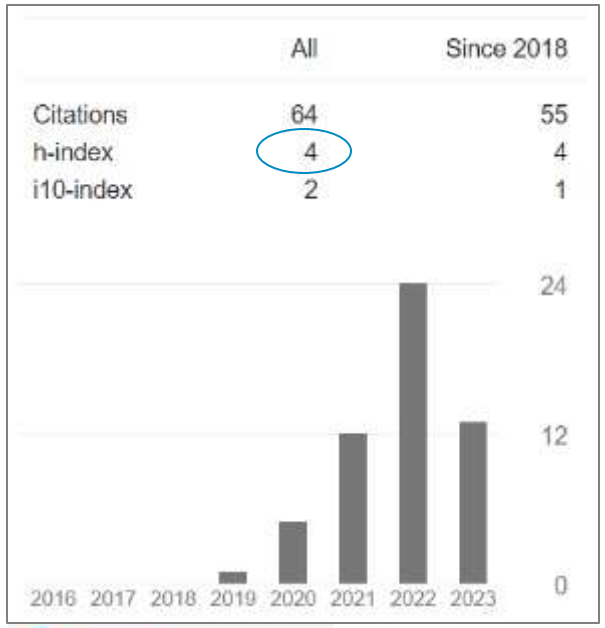




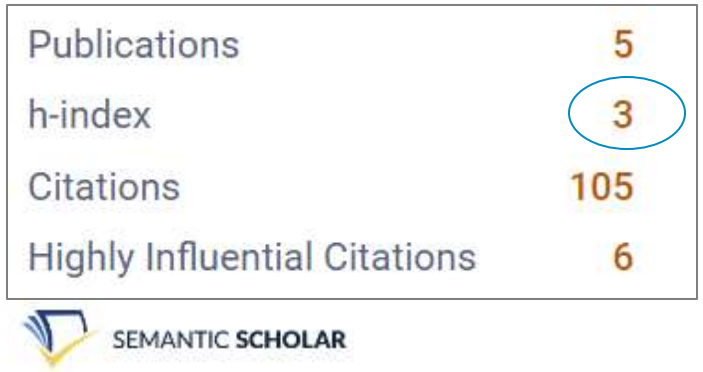
Human-Centered AI, White Paper. June 2021. CMU's Softw are Engineering Institute. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

Carol J. Smith. Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. https://arxiv.org/abs/1910.0351
Checklist and Agreement - Downloadable PDF: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620
Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. https://www.diu.mil/responsible-ai-guidelines

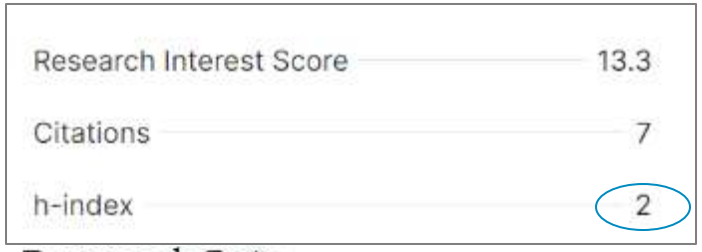
Publishing for an Academic Footprint



Google Scholar



SEMANTIC SCHOLAR



ResearchGate

Articles	Citation numbers	
1	33	
2	30	
3	20	
4	15	
5	7	
6	6	= h-index
7	5	
8	4	

Univ of Waterloo: https://subjectguides.uwaterloo.ca/calculate-academic-footprint/YourHIndex

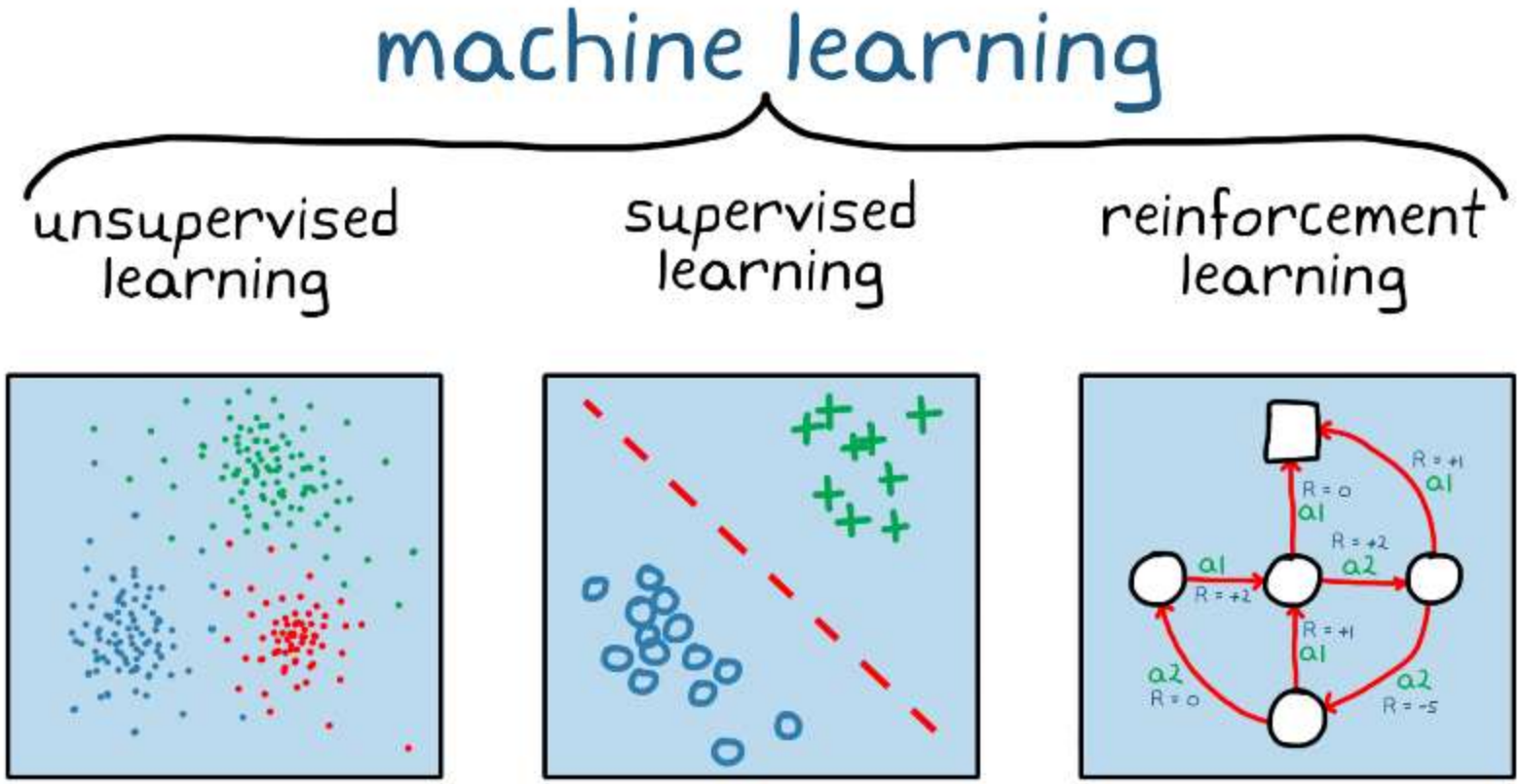
Making Responsible Artificial Intelligence

Carnegie Mellon University
Software Engineering Institute

AI systems can
- recognize patterns
- create predictions
- make decisions, and/or
- generate new content

without being explicitly programmed

Artificial Intelligence

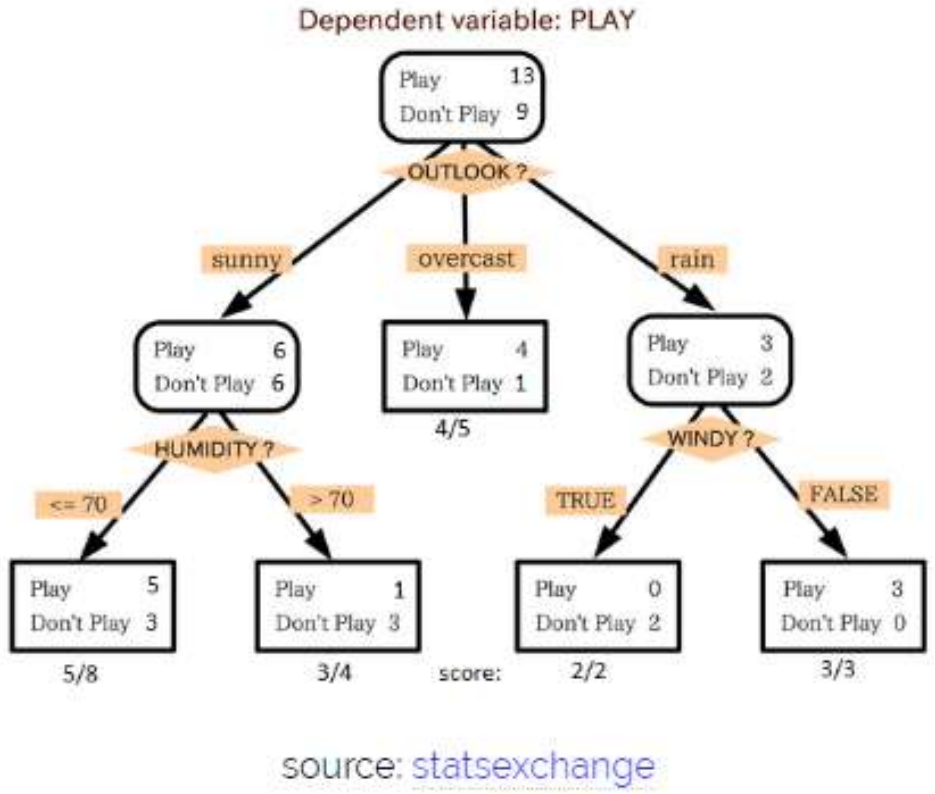


Deep Learning, Neural Networks, Large Language Models (LLMs)

AI / Machine Learning

- Algorithms
  - math + programming
- Model (AI)
  - Algorithms + data

- Know ONLY what taught
- Control ONLY given control of
- Aware of nuances
  - can continue to learn



https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/

Taxonomies and Ontologies coming to life

(NOT like humans learn)

AI is NOT sentient

Not unknowable

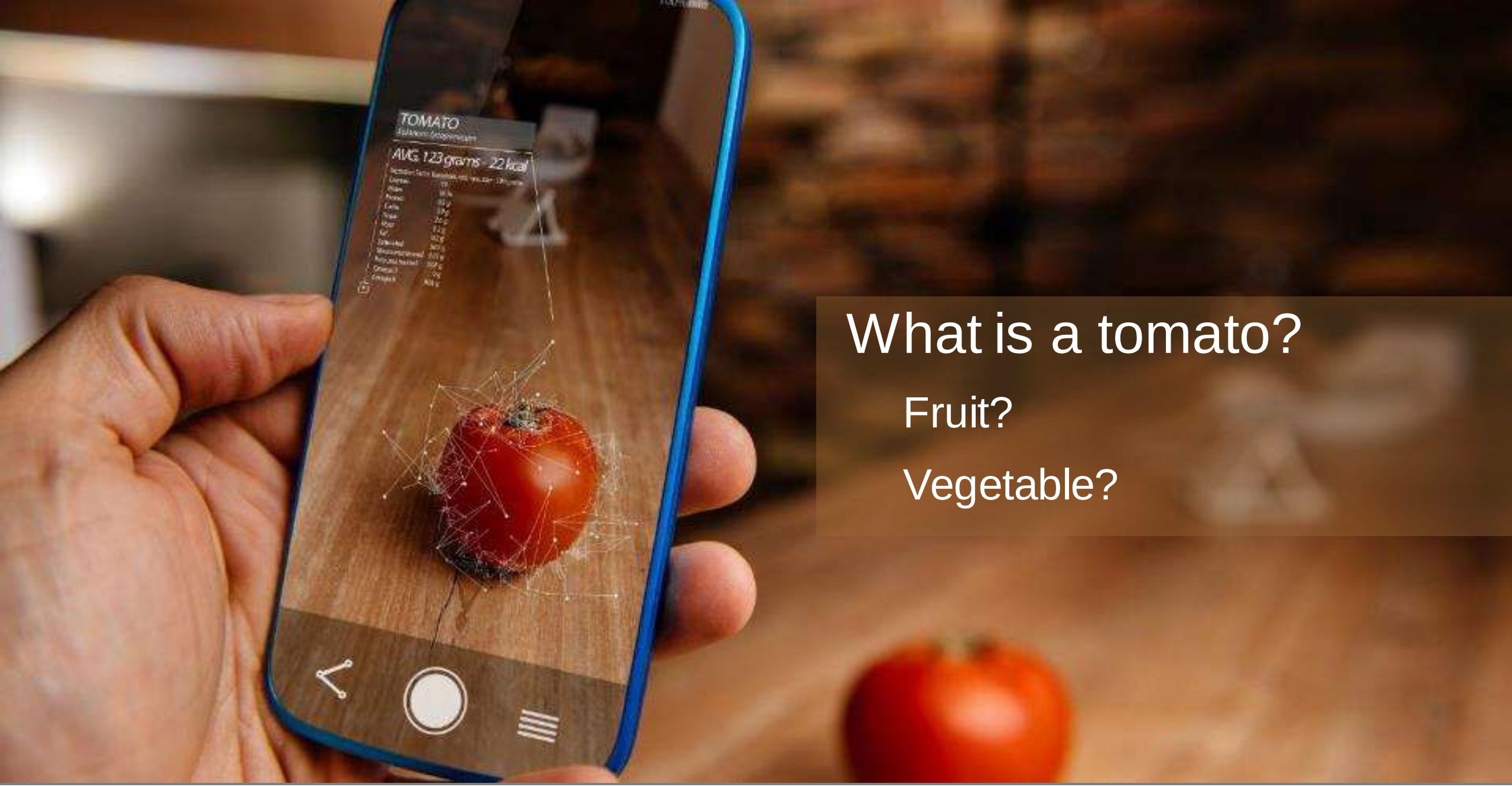
Never Enough Time

- Physician: ~90 hours reading a week*
- AI could bring that information to the physician
- Enabling more evidence-based decisions

Alper, Brian S. et al. “How Much Effort Is Needed to Keep up with the Literature Relevant for Primary Care?” Journal of the Medical Library Association 92.4 (2004): 429–437. Print.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC521514/

Transfer human concepts and relationships

Photo by sunlightfoundation
https://www.flickr.com/photos/sunlightfoundation/2385174105



What is a tomato?

Fruit?

Vegetable?

Computer Vision - Image Recognition

Carnegie Mellon University Software Engineering Institute

Train set

Data encountered

Use case courtesy of Dr. Eric Heim, CMU Software Engineering Institute
https://resources.sei.cmu.edu/library/author.cfm?authorid=542374

Only know what taught

Carnegie Mellon University Software Engineering Institute

Train set

Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

Carnegie Mellon University
Software Engineering Institute

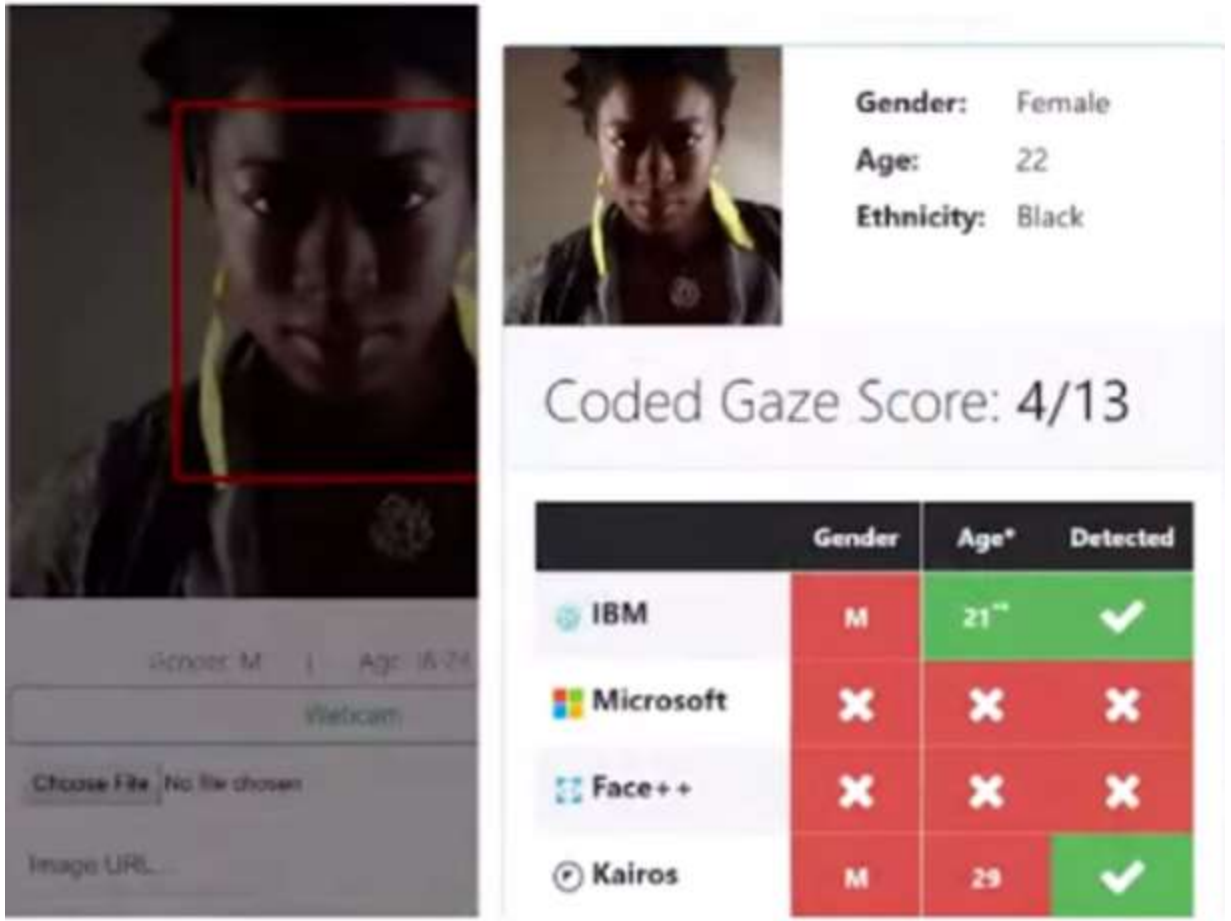
"Data is a function of our history…
The past dwells within…
Showing us the inequalities
that have always been there."

Joy Buolamwini, Algorithmic Justice League
Coded Gaze
Movie: Coded Bias on Netflix

Photo: Joy Buolamwini on The Open Mind: Algorithmic Justice. Jan 12, 2019. https://www.youtube.com/watch?v=hwHnXdoSSFY

Data is biased

- Not inherently neutral.
- Reflect priorities, preferences, and prejudices of people making them



Gender Shades Project by Joy Buolamwini. MIT Media Lab. https://www.media.mit.edu/projects/gender-shades/overview/

38

Algorithms of Oppression, Dr. Safiya Noble

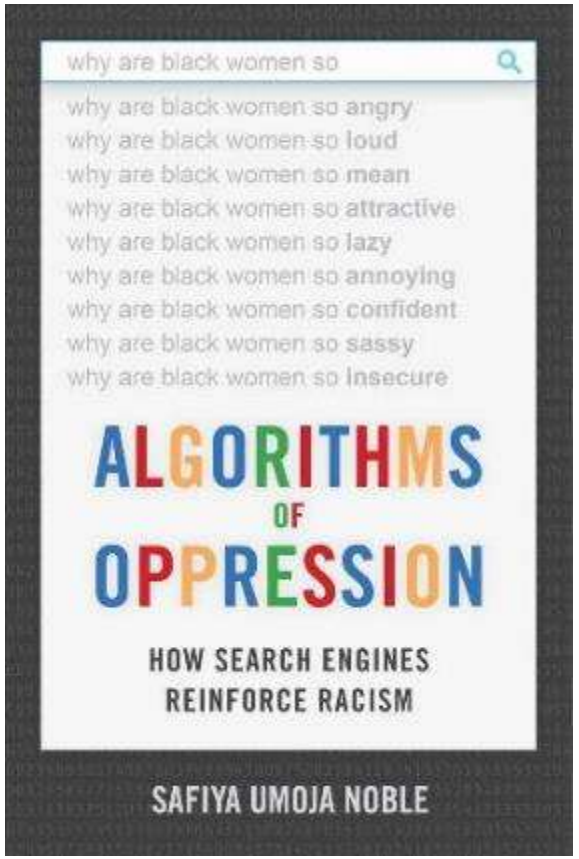


Photo from article: Google This: Algorithmic Oppression. ACLU News & Commentary. February 19, 2021. https://www.aclu.org/news/privacy-technology/google-this-algorithmic-oppression/

39

Carnegie Mellon University Software Engineering Institute

Responsible,

Intentional

Design

Just because you can, doesn't mean you should.

Conversations for Understanding

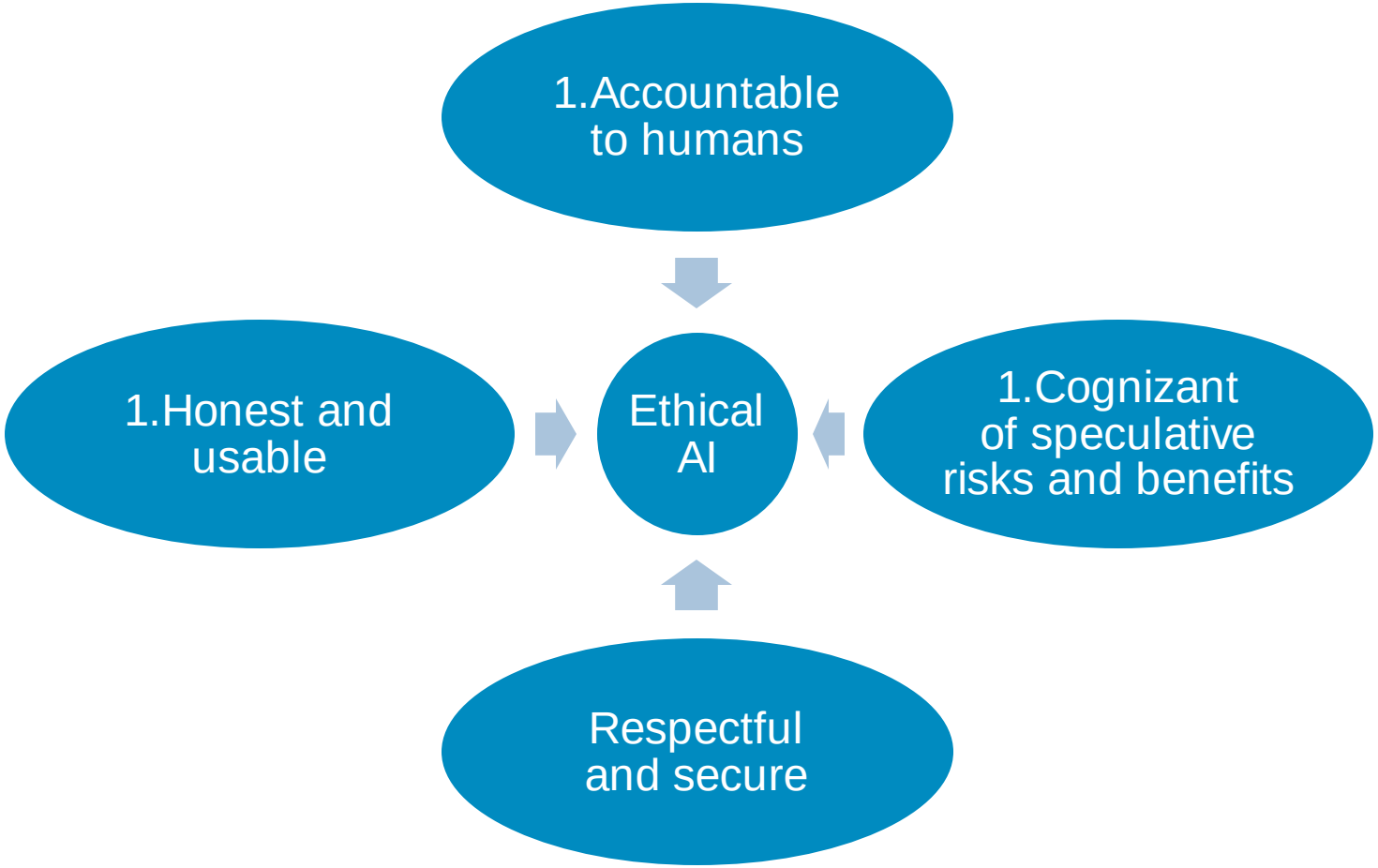
Difficult Topics

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

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

Photo by Pam On Unsplash

UX Framework for Designing Trustworthy AI



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.
https://insights.sei.cmu.edu/sei_blog/2020/03/designing-trustworthy-ai-for-human-machine-teaming.html

RightStaff Scenario

Carnegie Mellon University Software Engineering Institute

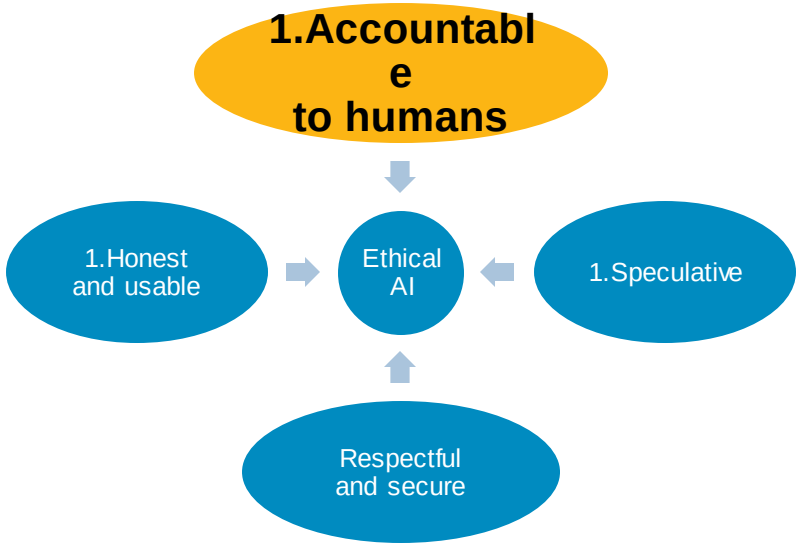
- AI shift scheduling system

- Users: Store managers of fast food restaurants

- Goals of RightStaff:
- Faster staffing decisions and scheduling
- Reduced bias of shift selection

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

Carnegie Mellon University Software Engineering Institute

Significant decisions

Carnegie Mellon University Software Engineering Institute

- Significant decisions made by the AI system will be
- explained
- able to be overridden
- appealable and reversible

- **RightStaff**
- Manager able to reschedule people as needed

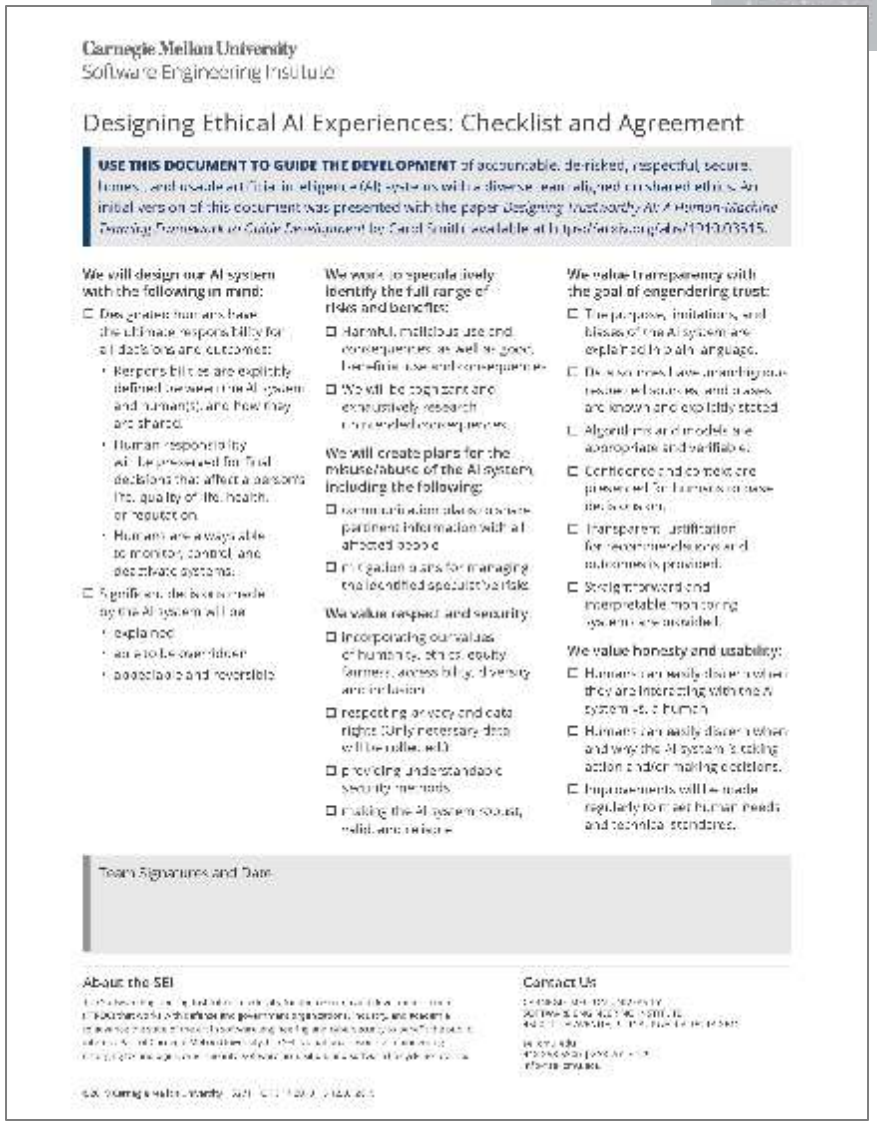
Responsibilities explicitly defined

- Between AI system and human(s)

- **RightStaff** (AI System *or Manager?*)
- Picks employees to schedule?
- Defines shifts?

Prompt conversations

- Pair Checklist with Technical Ethics
  - Bridges gap between "do no harm" and reality
- Reduce risk and unwanted bias
- Support inspection and mitigation planning



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, [illegible] 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 for AI-processed fiscal decisions that affect a person's (e.g. 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
  - possible and reversible

We want to specifically identify the full range of risks and benefits:

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

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

- ☐ [illegible] patient information with all affected people
- ☐ [illegible] plans for managing the identified specialist risks

We value respect and security:

- ☐ Incorporating our values of humanity, ethics, equity, fairness, access to [illegible] diversity and inclusion
- ☐ respecting privacy and data rights (only necessary data will be collected)
- ☐ providing understandable security methods
- ☐ making the AI system sound, valid, and robust

We value transparency with the goal of engendering trust:

- ☐ The purpose, limitations, and biases of the AI system are explained in plain language.
- ☐ Decisions, [illegible] and biases are known and explicitly stated
- ☐ Algorithms and models are appropriate and verifiable.
- ☐ Confidence and context are presented for humans to use decisions
- ☐ Transparent justification for recommendations and outcomes provided.
- ☐ Strong forward and interpretable monitoring system 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

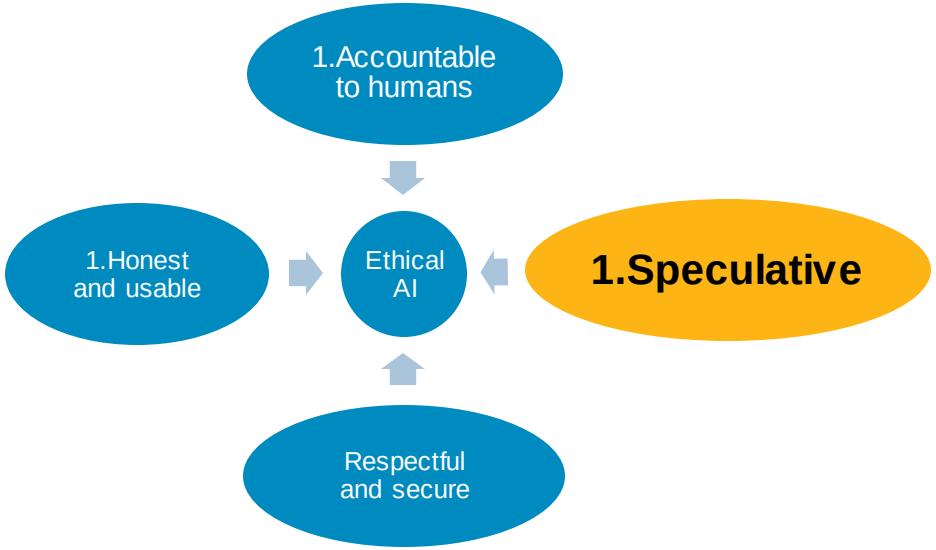
About the SEI

Contact Us

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

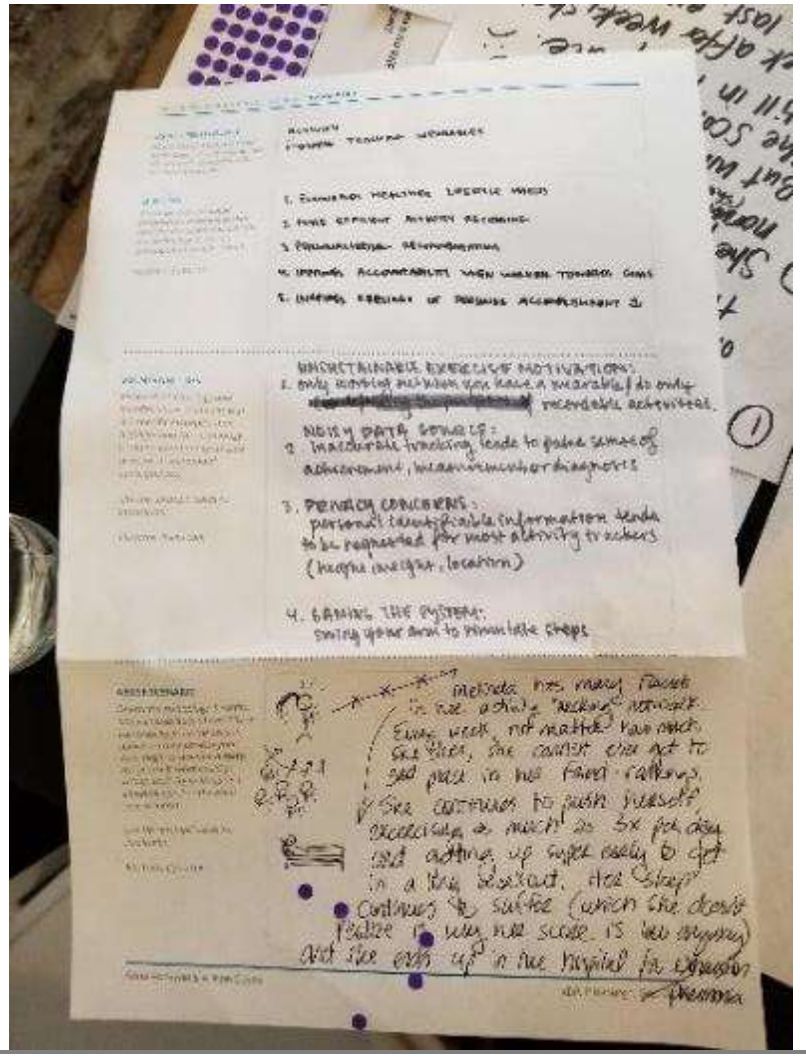
Cognizant of Speculative Risks and Benefits

- Identify full range of
  - Harmful, malicious use, as well as good, beneficial use
  - Unwanted/unintended consequences



Prevent bias - speculate about misuse and abuse

- At earliest moments (before coding)
- Activate curiosity
- Perspective of people in frequently marginalized groups



Template by Anna Abovyan & Allison Cosby, IxDA Pittsburgh, Sep 2019

Reward team members for finding ethics bugs

Bias Scenario

- RightStaff begins prioritizing people with easier schedules
- Managers approve these schedules, reinforcing bias
- People who were previously discriminated against are *still* discriminated against

Abuse Scenario

- Managers want to avoid providing benefits to employees
- RightStaff adjusted to ensure that no one has full shifts
- No regular staff can keep benefits

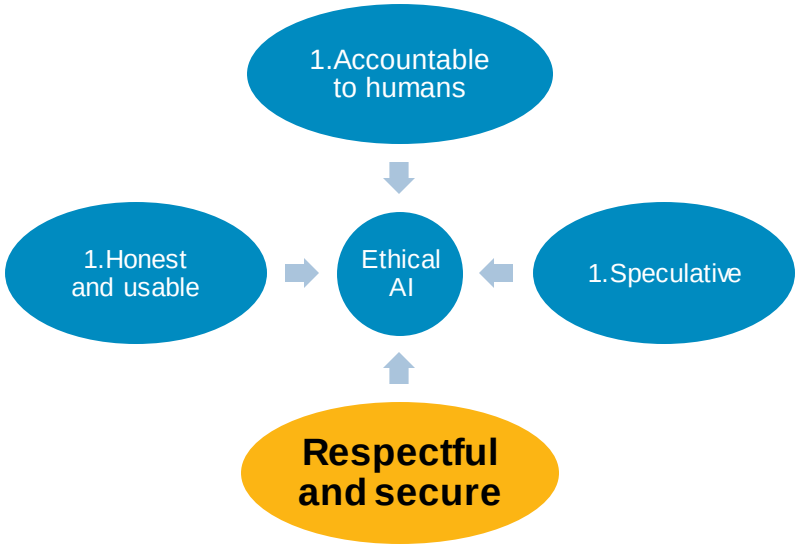
- What else?

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



Respectful and Secure

Carnegie Mellon University Software Engineering Institute

- **RightStaff**
- Who has visibility to reasons for changing schedules?
- How is that information used?
- How is PII* of employees protected?

*PII is Personally Identifiable Information (social security number, address, etc.)

Honest and Usable

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

Fair: Identify bias in data

Carnegie Mellon University Software Engineering Institute

- Show awareness of known and desirable bias
- Acknowledge and communicate about unwanted issues

RightStaff

- System built to reduce the known bias in existing data
- Make it easy to report bias (or prevent it)

AI has great potential, develop with caution

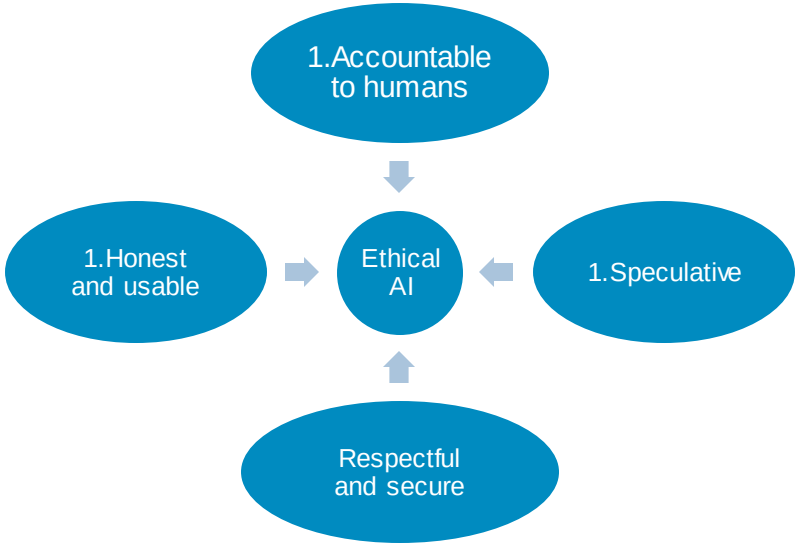
Carnegie Mellon University Software Engineering Institute

Future AI's may be trusted to substitute human cognition and abilities.

"AI will ensure appropriate human judgement and not replace it" - DIB

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

- Empower diverse teams, inclusive environments
- Adopt technical ethics
- Encourage deep conversations (Checklist)
- Activate curiosity; be speculative; imaginative



Carnegie Mellon University Software Engineering Institute

Evangelize for human values

Ethical. Transparent. Fair.

Advice

Carnegie Mellon University
Software Engineering Institute

Carnegie Mellon University Software Engineering Institute

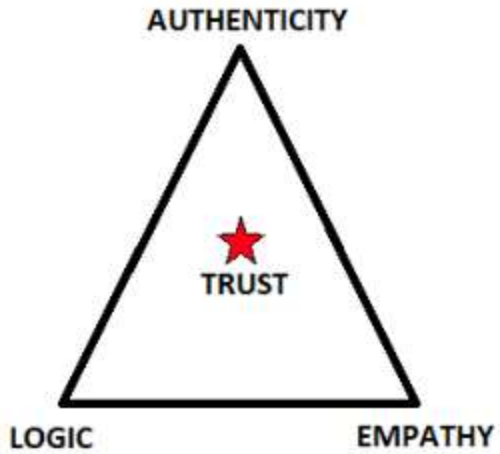
Make Relationships.

Find people who believe in you - keep in touch with them.

And…

- Trust your instincts
- Don’t plan too far out
- Learn what you can
- Ask for feedback
- Work is long - look for fun, interesting work

Authenticity, Empathy, Logic



Understand where you [and others] wobble
- Frances Frei, Harvard Business School professor

Frances Frei, Harvard Business School, https://blog.ted.com/how-to-rebuild-trust-frances-frei-speaks-at-ted2018/

Build a manager Voltron

Carnegie Mellon University Software Engineering Institute

Develop diverse manager crew/coaching network
Look for people who:

- push you out of your comfort zone
- have different levels of experience than you
- are from a different background
- are good at things you're terrible at

When your manager isn't supporting you, build a Voltron by Lara Hogan Originally posted Jan 4, 2018
https://larahogan.me/blog/manager-voltron/

Your Responsibility

- Question status quo
- Stand for ethics – do the right thing
- Talk about bias
- Be inclusive and kind

Find your priorities

Carnegie Mellon University Software Engineering Institute

- What you can put up with and what you cannot.
- Invest in experiences
- Learn what you can for as long as you can

- Not learning, being challenged?
  - Change is ok - need to be able to tell the story
  - Look/ask for new opportunities
  - Go and try something new

Q & A Time!

Carnegie Mellon University
Software Engineering Institute