World Usability Day
UXPA Minneapolis, November 2021

# Navigating the Complexity of Trust

Carol J. Smith
Sr. Research Scientist - Human-Machine Interaction, CMU's SEI
Adjunct Instructor, CMU's Human-Computer Interaction Institute

Twitter: @carologic @sei_etc

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

# Copyright Statement

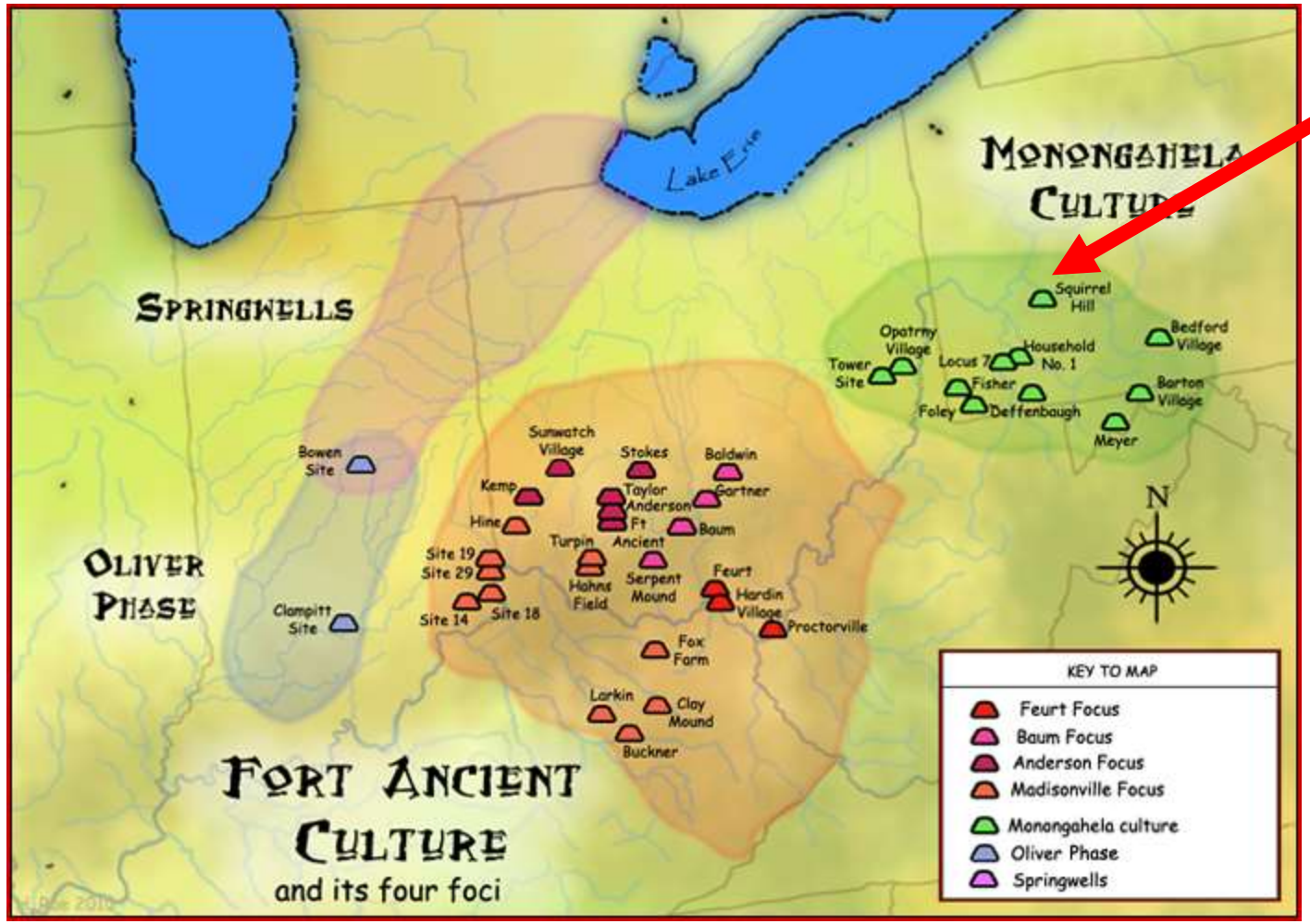


This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.



DM21-1037

# Acknowledging the Land I Speak On




Land of Monongahela, Adena and Hopewell Nations;
Seneca, Lenape and Shawnee lands;
Osage, Delaware and Iroquois lands.

Now known as Pittsburgh, PA, USA.

Map by Herb Roe via Wikipedia https://en.wikipedia.org/wiki/Monongahela_culture

# What is Trust?

# Complex, Transient, and Personal

# Contradictions



Jonathan Rotner, Ron Hodge and Lura Danley. 2020. AI Fails and How We can Learn from Them. The MITRE Corporation. July 2020. Case number 20-1365. https://sites.mitre.org/aifails/failure-to-launch/

**Trust is having enough confidence* in positive outcomes, that a person gives control of something significant to them, to another party.**

*Confidence is based on evidence of benevolence, integrity, and capability.

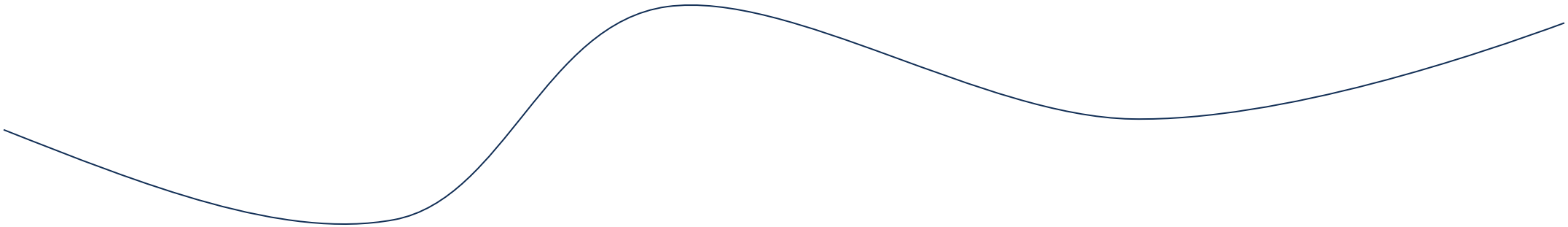
# Trust Involves…

- Belief and understanding
- Dependency and choice
- Context and privacy
- Perception and awareness
- Evidence and knowledge
- Emotion and respect

Jonathan Rotner, Ron Hodge and Lura Danley. 2020. AI Fails and How We can Learn from Them. The MITRE Corporation. July 2020. Case number 20-1365. https://sites.mitre.org/aifails/failure-to-launch/

# Appropriate Trust

As **variations occur** in context,
and confidence in evidence* changes,
there is a corresponding adjustment
in the **level of trust**.



*of benevolence, integrity, and capability,

Building on Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. IUI 2017 (March 2017), 307-317. DOI: http://dx.doi.org/10.1145/3025171.3025219

# 100% Trust?

# What is Appropriate?

Can there be too much trust?

What is necessary?

How do we communicate what is appropriate?

# Optimal Trust

"Unnecessarily high trust in AI
may have catastrophic consequences,
especially in life-critical applications…

Optimal trust in which both humans
and AI each have **some level of skepticism**
regarding the other's decisions
since **both are capable of making mistakes**."

Onur Asan, Alparslan Emrah Bayrak and Avishek Choudhury. 2020. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res (2020), Vol. 22, 6:e15154. URL: https://www.jmir.org/2020/6/e15154 DOI: https://doi.org/10.2196/15154

# Semi-Autonomous Vehicles


Tesla Autopilot in Heavy LA Traffic by Scott Kubo https://youtu.be/m3-QzTFxoUg?t=14

# Automation Bias

Propensity for humans to **favor suggestions** from automated decision-making systems and to **ignore contradictory information** made without automation, even if it is correct.

Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA 2004-6313. AIAA 1st Intelligent Systems Technical Conference. (September 2004). DOI: https://doi.org/10.2514/6.2004-6313

# Trust is a Continuum

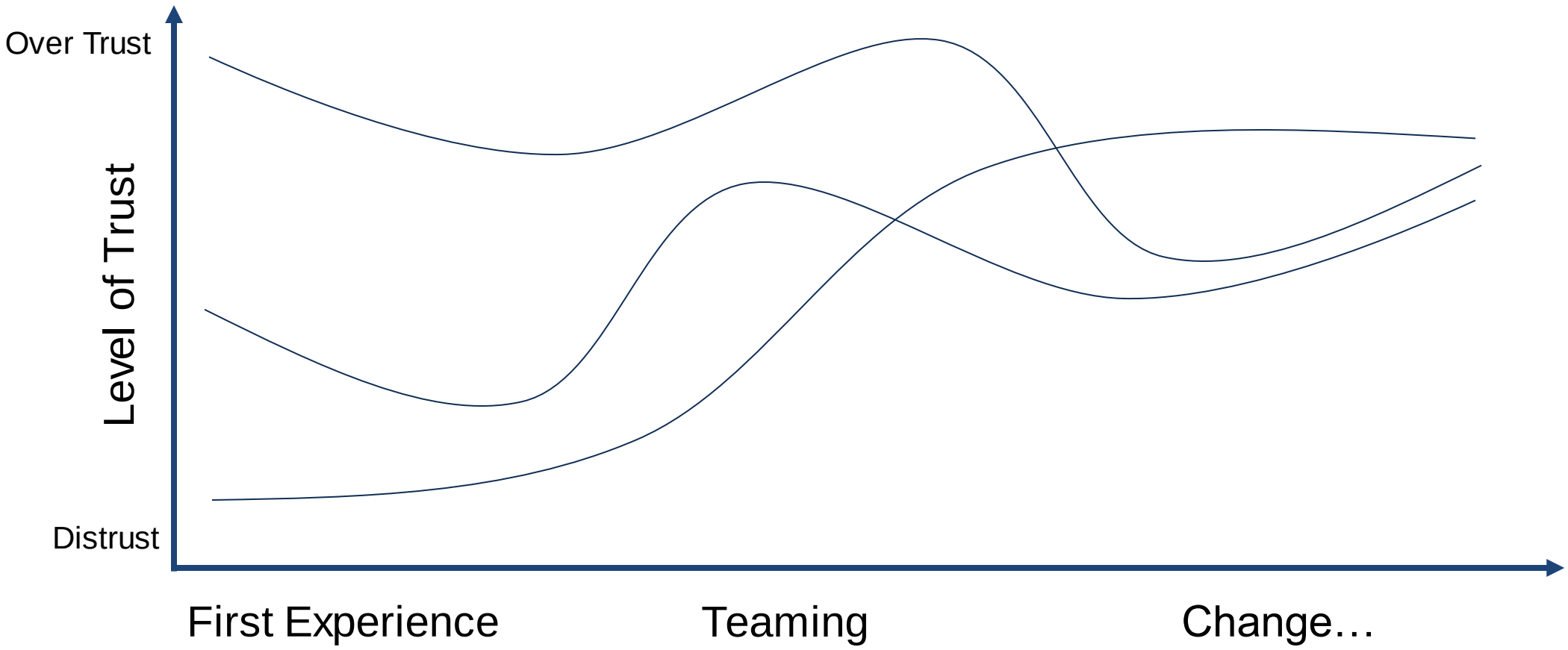
**Distrust**
Trust falling short of system capabilities - may lead to disuse.

**Calibrated Trust**
Trust matches system capabilities leading to appropriate use.

**Over Trust**
Trust exceeding system capabilities - may lead to misuse

Bobbie Seppelt and John Lee. 2012. Human Factors and Ergonomics in Automation Design. In Handbook of Human Factors and Ergonomics (Fourth Edition) Chapter 59. Wiley. DOI: https://doi.org/10.1002/9781118131350.ch59

# Trust Changes Over Time




Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. IUI 2017 (March 2017), 307-317. DOI: http://dx.doi.org/10.1145/3025171.3025219

# Change Increases or Decreases Trust

## Event-Driven

- Response to an interaction, transaction, service, or event

## Time-Driven

- Response to periodic evidence (observations, recommendations)
- Lack of evidence can decay trust

Jia Guo and Ing-Ray Chen. 2015. A Classification of Trust Computation Models for Service-Oriented Internet of Things Systems. 2015 IEEE International Conference on Services Computing (2015), 324-331. DOI: https://doi.org/10.1109/SCC.2015.52

Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. IUI 2017 (March 2017), 307-317. DOI: http://dx.doi.org/10.1145/3025171.3025219

# Change is Constant

# Awareness of System Capabilities

Transparency and usability of system

Understanding of conditions, constraints

Gain awareness thru experience

- Length of time
- Quality of experience

# Additional Trust / Distrust Factors

Institutional, management

Social and relational

Previous experiences


# Evidence:

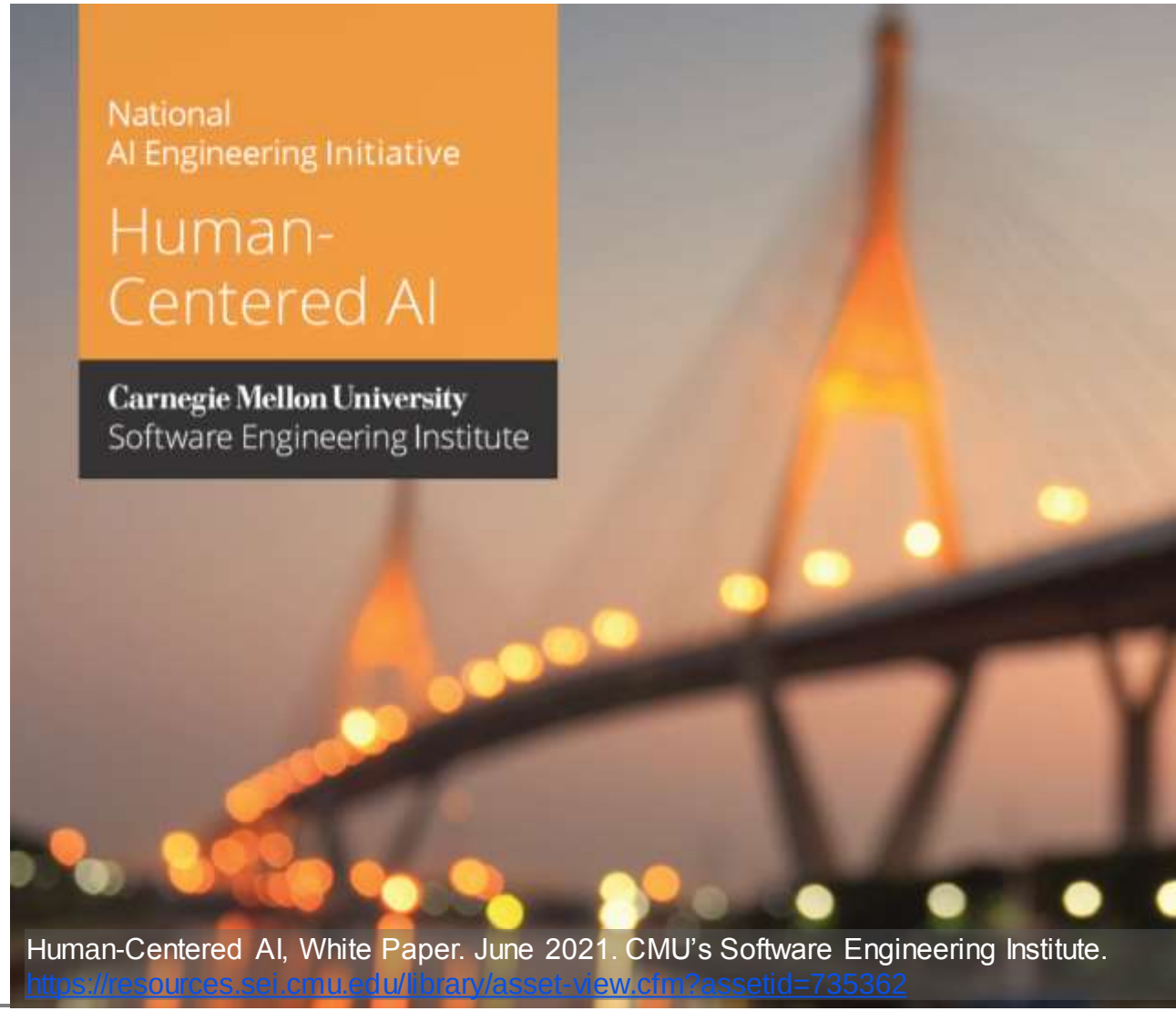
- Benevolence?
- Integrity?
- Capability?

# Supporting Appropriate Trust

# Design to work with, and for, people

## Minimize unintended consequences

- Research to understand context of use
- Design for purpose: Systems – not just tasks
- Test prototypes/products in environment




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


# Speculation keeps people safe

# Speculate and Design for the Worst Case

Be speculative about the worst case

- Don’t assume only average cases
- Probabilities about what will happen in the future can't be verified
- Don’t require unsupportable risk assessments.

N. G. Leveson. 2017. The Therac-25: 30 Years Later. In Computer, vol. 50, no. 11, (November 2017), 8-11. DOI: 10.1109/MC.2017.4041349

# Activate Curiosity

UX research methods and activities to activate curiosity:

- Abusability Testing (Dan Brown)
- "Black Mirror" Episodes (Casey Fiesler)
  (inspired by British dystopian sci-fi tv series of same name)

Speculate about system misuse and abuse

- What are potential unintended/unwanted consequences?

# 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

# New uncomfortable work

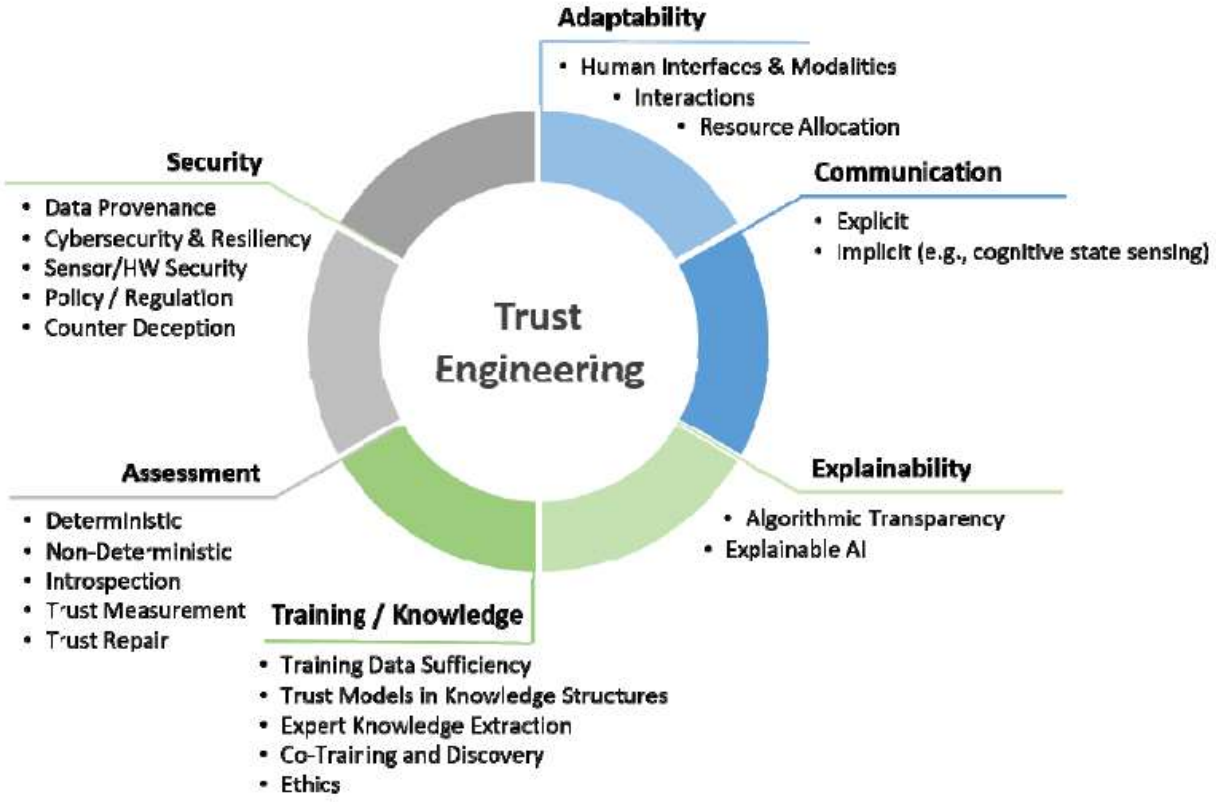
## “*Be uncomfortable*”

- Laura Kalbag

Ethical design is not superficial.

# Designing for Trust

# Critical Design Components




- Security
- Adaptability
- Communication
- Explainability
- Training/Knowledge
- Assessment

Neta Ezer, Sylvain Bruni, Yang Cai, Sam J. Hepenstal, Christopher A. Miller, and Dylan D. Schmorrow. 2019. Trust Engineering for Human-AI Teams. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63, no. 1 (November 2019): 322–26. https://doi.org/10.1177/1071181319631264.

# Safe Experiences

Actions to get into or maintain
a **safe state** should be **easy** to do.

Actions that can lead to
an **unsafe state** (hazard) should be **hard** to do.

Don’t rely on operators to detect errors
and recover before an accident – it isn't realistic.

N. G. Leveson. 2017. The Therac-25: 30 Years Later. In Computer, vol. 50, no. 11, (November 2017), 8-11. DOI: 10.1109/MC.2017.4041349
N. Leveson. 1995. Safeware: System Safety and Computers, Addison Wesley (1995).

# Take Pause

"In our enthusiasm
to provide measurements,
we should not attempt to measure
the unmeasurable."

**People believe a "calculated number more than actual experience."**

N. G. Leveson. 2017. The Therac-25: 30 Years Later. In Computer, vol. 50, no. 11, (November 2017), 8-11. DOI: 10.1109/MC.2017.4041349

# Errors and Information

Create protection against errors

Provide self-checks / error-detection / error-handling

Identify trends and behaviors that increase risk

**“Data—“big”**
**or not**
**—isn't the same as information.”**

N. G. Leveson. 2017. The Therac-25: 30 Years Later. In Computer, vol. 50, no. 11, (November 2017), 8-11. DOI: 10.1109/MC.2017.4041349

# Make Systems Effective Team Players

Easy to direct

- How observable is its behavior?
- How easily and efficiently allows itself to be directed?
- Even (or especially) during busy, novel episodes?

S. W. A. Dekker and D. D. Woods. 2002. MABA-MABA or Abracadabra? Progress on Human–Automation Co-ordination. Cognition Tech Work 4, (2002) 240–244. DOI: https://doi.org/10.1007/s101110200022 Note: MABA-MABA (Men-Are-Better-At/Machines-Are-Better-At lists)

# Capitalize on Human Strengths

Humans are better at:

- Perceiving patterns
- Improvising and using flexible procedures
- Recalling relevant facts at the appropriate time
- Reasoning inductively
- Exercising judgment

Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA 2004-6313. AIAA 1st Intelligent Systems Technical Conference. (September 2004). DOI: https://doi.org/10.2514/6.2004-6313

# What Changes Across Time Cycles?

Length of interactions

- Short and hectic
- Longer, cyclical

Iterative

Require clear communication, negotiation,
and coordination.


How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21) – Video https://www.designforcontext.com/ia-shaping-human-ai-collaboration

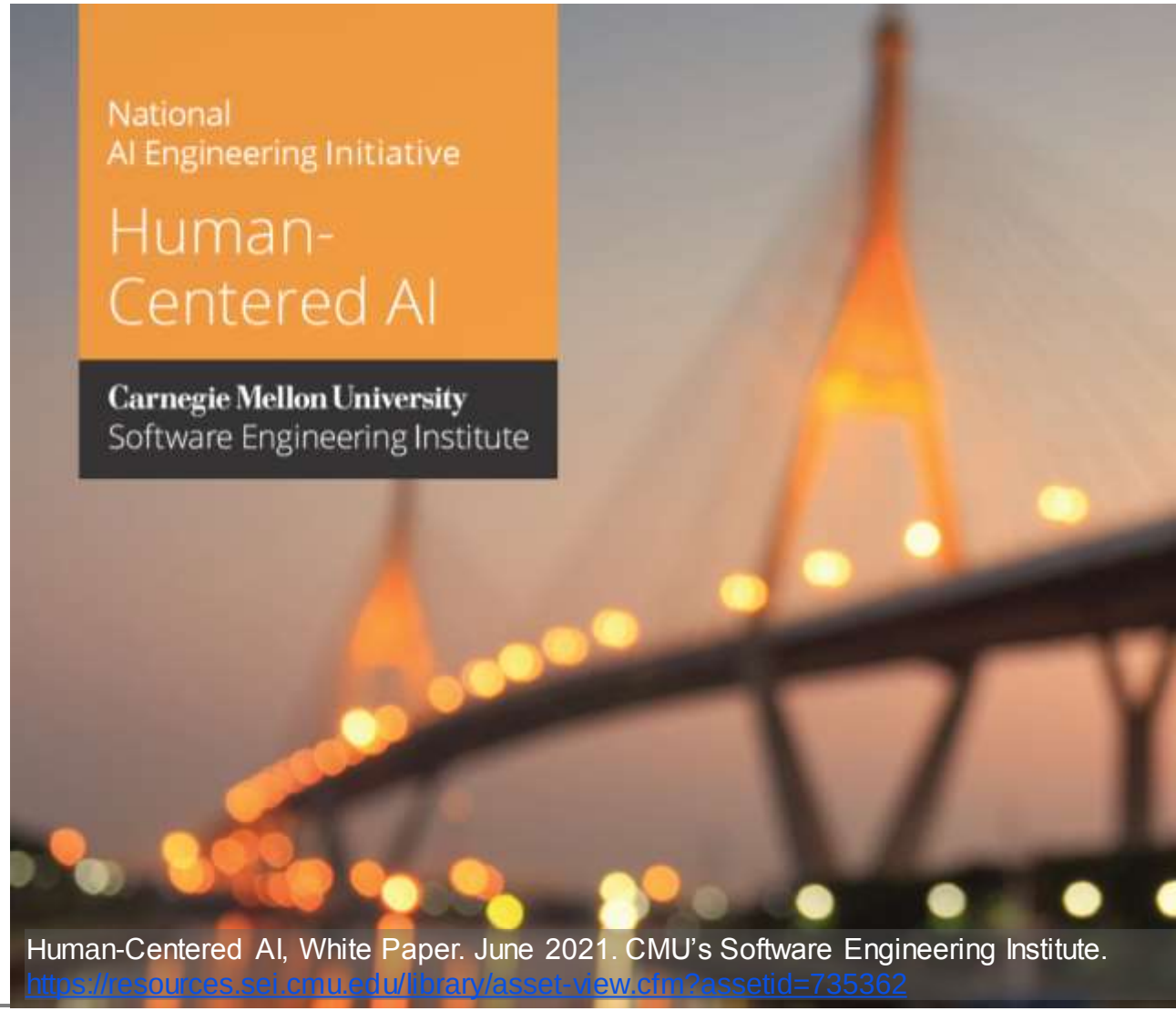
# Transparency

"Explainability" isn't magic.

Transparency isn't clarity.

System limitations

Boundaries
and unfamiliar scenarios



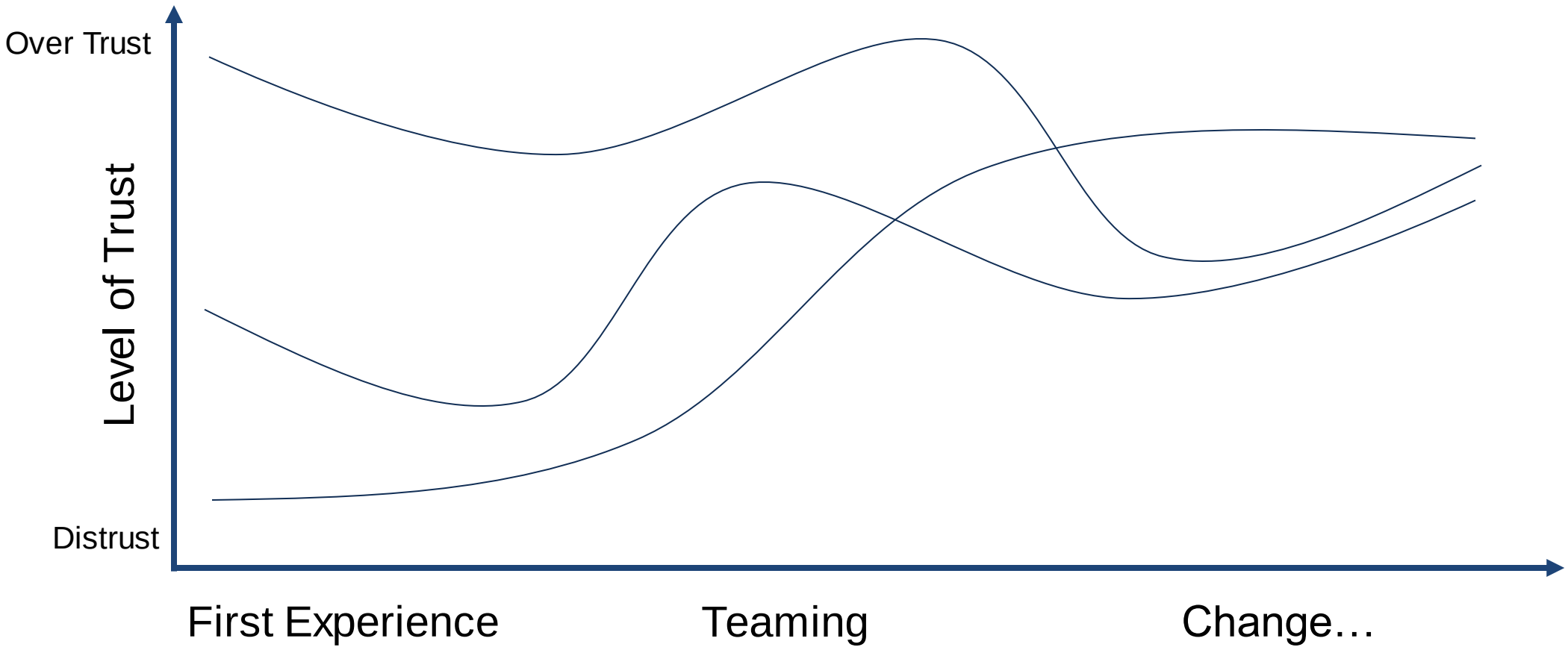

Human-Centered AI, White Paper. June 2021. CMU's Software Engineering Institute. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

# Appropriate Trust

- Understand context and test in context
- Design for purpose: Systems
- Provide understandable evidence
- Complement human strengths
- Provide control to people

Jonathan Rotner, Ron Hodge and Lura Danley. 2020. AI Fails and How We can Learn from Them. The MITRE Corporation. July 2020. Case number 20-1365. https://sites.mitre.org/aifails/failure-to-launch/

# Design for Appropriate Trust




Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. IUI 2017 (March 2017), 307-317. DOI: http://dx.doi.org/10.1145/3025171.3025219

Carol J. Smith

Twitter: @carologic

LinkedIn: https://www.linkedin.com/in/**caroljsmith**/

CMU's Software Engineering Institute,
AI Division

Twitter: @sei_etc

# Resources


Denise Rousseau, Sim Sitkin, Ronald Burt, and Colin Camerer. (1998). Not So Different After All: A Cross-discipline View of Trust. July 1988. Academy of Management Review. 23. 10.5465/AMR.1998.926617. DOI: 10.5465/AMR.1998.926617

Bobbie Seppelt and John Lee. 2012. Human Factors and Ergonomics in Automation Design. In Handbook of Human Factors and Ergonomics (Fourth Edition) Chapter 59. Wiley. DOI: https://doi.org/10.1002/9781118131350.ch59

Human-Centered AI, White Paper. June 2021. Carnegie Mellon University's Software Engineering Institute. Contributors: Hollen Barmer, Rachel Dzombak, Matt Gaston, Jay Palat, Frank Redner, Carol J. Smith. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

Jia Guo and Ing-Ray Chen. 2015. A Classification of Trust Computation Models for Service-Oriented Internet of Things Systems. 2015 IEEE International Conference on Services Computing (2015), 324-331. DOI: https://doi.org/10.1109/SCC.2015.52

Jonathan Rotner, Ron Hodge and Lura Danley. 2020. AI Fails and How We can Learn from Them. The MITRE Corporation. July 2020. Case number 20-1365. https://sites.mitre.org/aifails/failure-to-launch/

Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. IUI 2017 (March 2017), 307-317. DOI: http://dx.doi.org/10.1145/3025171.3025219

Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA 2004-6313. AIAA 1st Intelligent Systems Technical Conference. (September 2004). DOI: https://doi.org/10.2514/6.2004-6313

Neta Ezer, Sylvain Bruni, Yang Cai, Sam J. Hepenstal, Christopher A. Miller, and Dylan D. Schmorrow. 2019. Trust Engineering for Human-AI Teams. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63, no. 1 (November 2019): 322–26. https://doi.org/10.1177/1071181319631264.

N. G. Leveson. 2017. The Therac-25: 30 Years Later. In Computer, vol. 50, no. 11, (November 2017), 8-11. DOI: 10.1109/MC.2017.4041349

N. Leveson. 1995. Safeware: System Safety and Computers, Addison Wesley (1995).

Onur Asan, Alparslan Emrah Bayrak and Avishek Choudhury. 2020. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res (2020),Vol. 22,6:e15154. URL: https://www.jmir.org/2020/6/e15154 DOI: https://doi.org/10.2196/15154

Rose Challenger, Chris W. Clegg and Craig Shepherd. 2013. Function allocation in complex systems: reframing an old problem. Ergonomics, 56:7 (2017) 1051-1069. DOI: 10.1080/00140139.2013.790482