Panel: Trustworthy AI for HSI

June 8, 2020, http://hsi2020.welcometohsi.org/panel/

*Mr. **Zachary D. Tudor**, CISSP, CISM, Associate Laboratory Director, National and Homeland Security Idaho National Laboratory*
*Prof. **Milos Manic, Panel Moderator**, Director, VCU Cybersecurity Center, Professor, Computer Science Dept., Virginia Commonwealth University; Affiliate, Idaho National Laboratory*
*Prof. **Jacek Ruminski, Panel Host**, Head of AI Bay, Head of Biomedical Engineering Department, Gdansk University of Technology*
*Prof. **Hideyuki Sawada**, Department of Applied Physics, School of Advanced Science and Engineering, Waseda University*
*Prof. **Pitoyo Hartono**, Department of Electrical and Electronics Engineering, School of Engineering, Chukyo University*

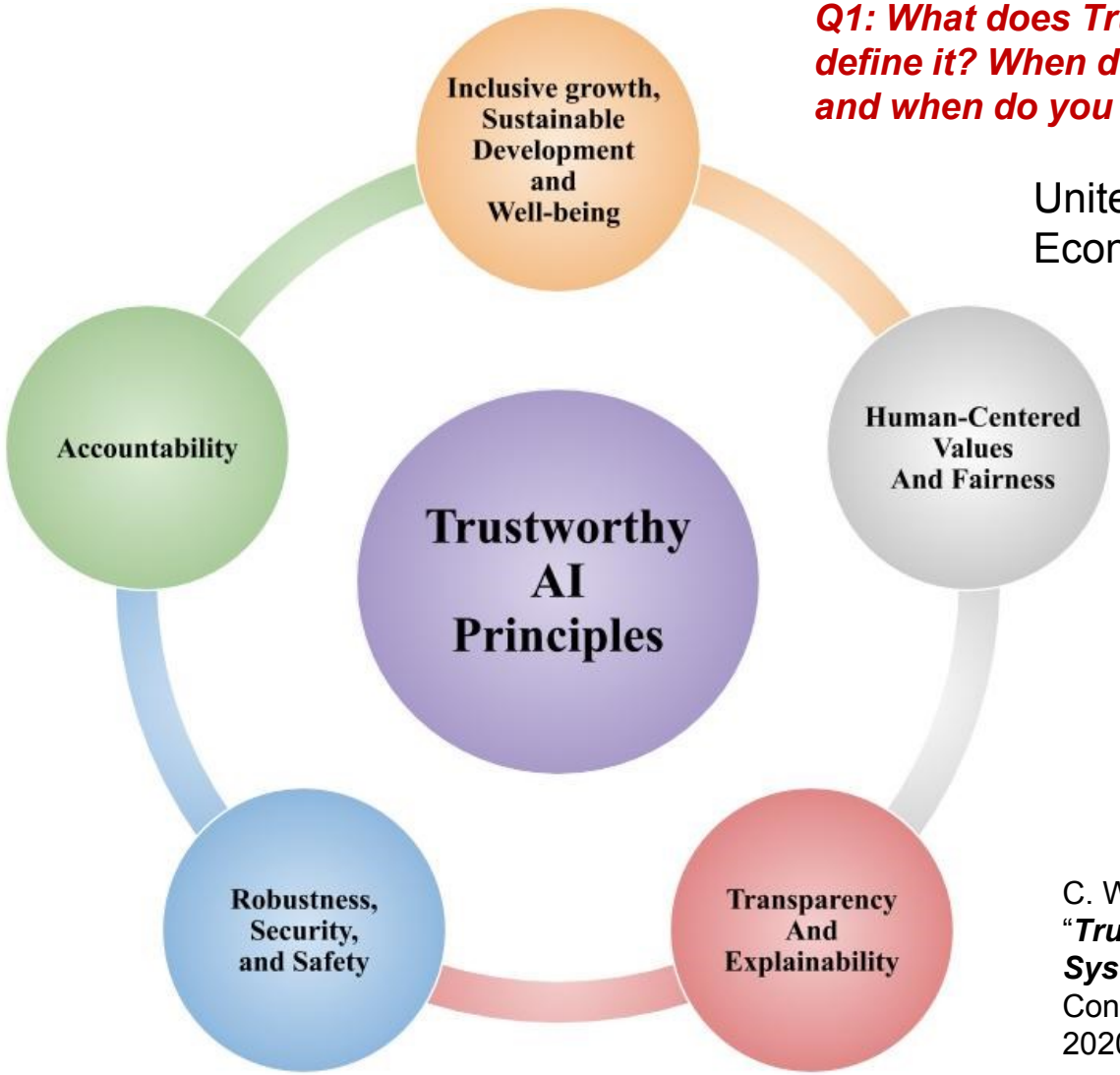
Trustworthy AI

Q1: What does Trustworthy AI mean to you, how would you define it?
When do you trust AI powered services/products and when do you not?
Why is Trust important?

- Definition (one of many…)
  - *Ethical principles* together with formal *AI system verification techniques* to define trustworthy AI, with the common goal of allowing people and societies to develop, deploy, and use AI systems without fear

- High-Level Expert Group on Artificial Intelligence (HLEGAI):
  - 'Striving towards Trustworthy AI concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle'

Trustworthy AI Principles



Q1: What does Trustworthy AI mean to you, how would you define it? When do you trust AI powered services/products and when do you not? Why is Trust important?

United States & OECD (Organization for Economic Cooperation and Development)

- Explainable AI
- Fair and Unbiased AI
- Privacy-Preserving AI
- Reliable/Verifiable AI

Crosscutting Areas

- Education and Workforce Development
- Policy, Governance, Ethics, and Privacy

C. Wickramasinghe, D. Marino, J. Grandio, and M. Manic, "***Trustworthy AI Development Guidelines for Human System Interaction***," in Proc. 13th International Conference on Human System Interaction, IEEE HSI 2020, Tokyo, Japan, June 6-8. 2020.

Q2: Which industrial sectors are a lesser candidate for AI support, if any?

How do you see the penetration of AI driven systems (government vs. private sector, power plants/grid vs. communications)?

Three laws of *robotics* (Asimov)

(Lower numbered laws supersede the higher numbered laws;
a 'zeroth law‘ later added)

Q3: AI-powered human system interaction are taking crucial role in many sectors, from smart factories to personal assistants (e.g. for the elderly).

Do we need a different trustworthy approach or maybe we can find some common rules (like Isaac Asimov's "Three Laws of Robotics")?

Law Zero:

- *(4th, added later) A robot may not injure humanity, or, through inaction, allow humanity to come to harm.*

Law One:

- *A robot may not injure a human being, or, through inaction, allow a human being to come to harm (unless this would violate a higher order law).*

Law Two:

- *A robot must obey orders given it by human beings, except where such orders would conflict with a higher order law (first law).*

Law Three:

- *A robot must protect its own existence as long as such protection does not conflict with a higher order law (first or second law).*

(Taken from http://www.robotics.utexas.edu/rrg/learn_more/history/#definition; https://en.wikipedia.org/wiki/Three_Laws_of_Robotics#First_Law_modified)

Trustworthy AI

Q5: AI augmentation vs. AI automation?

- **Trust:** predictable behavior, even in the presence of uncertainty.
- Two main components:
  - Intentions
  - Competence

- **Trustworthy AI**: combination of diverse research areas on AI systems:
  - Fairness, robustness, explainability, accountability, verifiability, transparency, and sustainability
  - Goals:
    - Identify factors which harm the human trust of AI systems
    - Introduce methods to improve human trust in AI systems

Fig. 5. Testbed

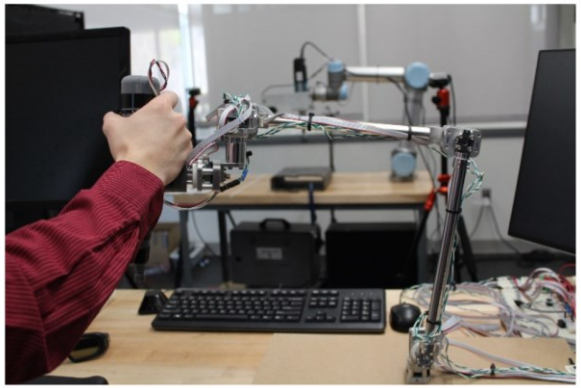


Fig. 6. Input Device

D. Marino, J. Grandio, C. Wickramasinghe, K. Schroeder, K. Bourne, M. Manic, "***AI Augmentation for Trustworthy AI: Augmented Robot Teleoperation***" in Proc. 13th International Conference on Human System Interaction, IEEE HSI 2020, Tokyo, Japan, June 6-8. 2020

Augmented AI for Trustworthy AI

Q5: AI augmentation vs. AI automation?

- Augmented AI: AI technologies working alongside humans
  - + Improve productivity, efficiency, quality of human activities, and enhance human-machine cognition
  - + Build trust
- Shared Autonomy
  - + Split tasks between AI and Humans
  - + High risk decisions made by humans
  - + Maintain *Accountability* in humans

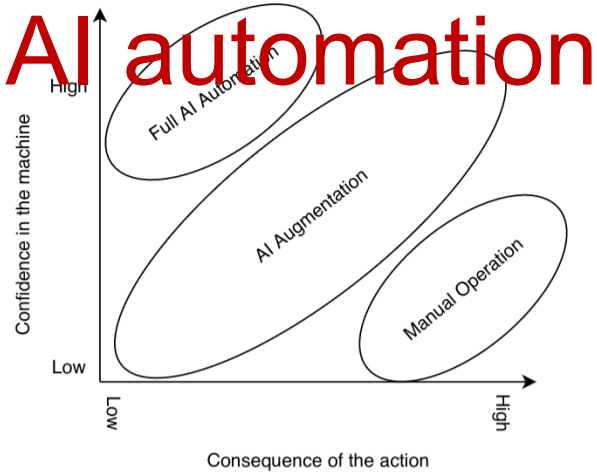


Fig. 2. Augmentation vs Automation

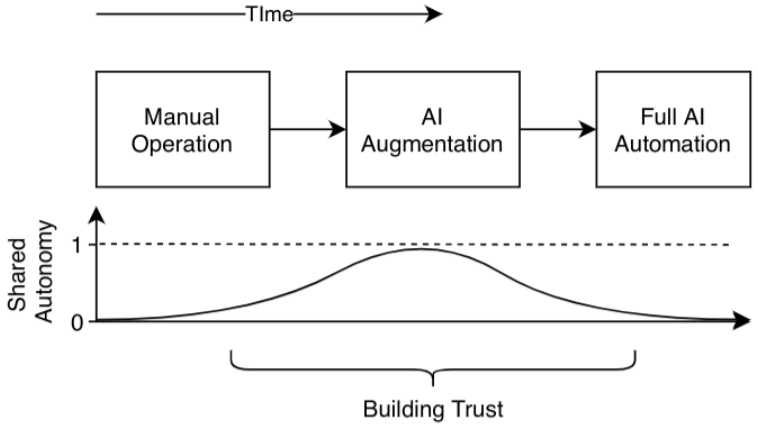


Fig. 3. Building trust

Q4: What is the future of AI-powered HSI?

Q6: What is a pathway towards improvement of human trust in human-AI interactions?