Trust and trustworthiness

- Tarek Abdelzaher
- Lori Clarke
- Charles Friedman
- Susan Graham
- Joshua Rubin
- Bill Scherlis
- Laurie Williams

CCC♪

"Trust is essential for CSLS" - Ed♪

Trust and trustworthiness

- **Definitions**
  - **Trustworthiness**
    - Attributes of a system (including its human participants)
      - E.g., flight controls in aircraft
  - **Trust**
    - Attitude of human operators and stakeholders with regard to a system
      - E.g., Tesla (over-trust); vaccinations (under-trust)

Trust → / Trustworthiness V	High	Low
High		*Under-trust*
Low	*Over-trust*	

  - **Participant**
    - Human in some role as part of a system: operator, user, etc.
    - E.g., MD, patient, RN, driver, student, teacher, data analyst
  - **Stakeholder**
    - Human and organizations affected by the actions of a system
    - Includes participants

Trust and trustworthiness

- **Trustworthiness attributes**
  - Categories
    - Security, safety, privacy, reliability, resilience, ethics
      - High dimensionality for each of these categories
    - Different kinds of assays (technique to make a judgment) for each
  - Evaluation practices and measures
    - For each attribute
      - We identify thresholds, tolerances, and norms – perfection unattainable
    - Support for human judgment of fitness or "degree of trustworthiness"
    - Authorities and certification
      - Systems *and* participants
- **Trust attributes**
  - Vary according to roles and backgrounds of participants and stakeholders
  - Influenced by extrinsic events
    - Accidents, changes in context
    - E.g., changing demographics in a city, changing student preparation
  - Influences by culture

Trust and trustworthiness

- **Influences on trustworthiness**
  - Assays for the individual **attributes**
    - Use of technical means to inform human judgment and certification
      - Light under lamppost vs. Lord Kelvin
    - Effectiveness: must thwart adversaries (insiders and external)
      - VW, Theranos, …
    - Modeling of functional and performance needs
    - Modeling of operating environment and scope of interactions with it
  - **Architecture** and **composition** among components – technical challenge
    - Governance as a primary influence on architecture and architecture evolution
  - Trustworthiness **designed into** the system
    - Influence on engineering model design, on "engineering data," and on tooling
  - Monitoring, logging, and **dynamic response**
  - Engineering the **participant experience** to enhance trustworthiness
    - E.g., avoiding password stickies
    - E.g., correct metaphors for policies and processes

- **Influences on trust**
  - **Explanation** and transparency
    - Explanation influenced by implicit knowledge of stakeholder
  - **Perspicuity of metaphors** presented to humans [participants and stakeholders]
  - **Experience** with the system over time
  - **Governance** – identified stakeholders; drives trustworthiness practices
  - **Business factors**: compliance, safe harbors, incentives (+/-) of players

Trust and trustworthiness

- **Domains and examples**
  - Health care
    - E.g., shortcuts and optimization in procedures, based on data
  - Cities
    - E.g., adaptive traffic management (AI Social Good)
  - Education
    - E.g., student customized course materials
  - ATC
    - E.g., flight route optimization (Platzer)

- **Tensions in a learning system**
  - Establishment of routine and best practices **vs.** Benefits of continual adaptation
    - Being forced "out of the groove"
  - Transformation of the roles for humans
    - Cf. "usable security and privacy"
    - Participants, operators, users, other stakeholders
  - Architecture designed for "degree of trustworthiness"
    - Vs. usual "software discontinuity"
  - Usability, "invisibility," and trust
    - E.g., is the new Bay Bridge safe in an earthquake?
  - Ethical tensions
    - E.g., perturbations in traffic, Facebook experiments, Peoria

Additional research topics

- **What are the trustworthiness attributes?**
  - How to assays for each?
  - What are measures?
  - What aspects are particularly challenging/special for CSLS
  - Relation with architectural decisions and composition
- **How can trust be measured, as it evolves over time?**
  - What are influences on trust?
  - How does "invisibility" or "embeddedness" influence trust?
- **New challenges in certification of trustworthiness?**
  - New influences
    - Rapid system-scale adaptation
    - Diffused governance (ULS)
    - Data quality (training) as an influence on system trustworthiness
  - Certification of ***process of adaptation*** as well as (in lieu of?) outcomes of adaptation
    - Emergence: What if we cannot easily evaluate the *results of adaptation* – but we have evaluated the adaptation mechanism?
    - How good are the data that inform that adaptation – what is evaluated here?
  - How to measure risk and "zone of uncertainty"
- **Is there a concept of "trust engineering" or at least "trust mgmt"?**
  - What are the ***principles of trust***?