Trustworthy
Cyber Social Learning Systems

Lori A. Clarke
College of Information and Computer Sciences
University of Massachusetts Amherst

No question that such systems "should be" trustworthy

- **Basis for important decision making**
- **Will impact health, well being, and safety of our citizenry**
  - micro decisions about individuals (e.g., medical care, education plans)
  - Macro decisions about best practices (e.g., standards of care, sustainable energy consumption)
- **Will have a tremendous economic impact**
  - On the cost of societal infrastructure
  - On individual companies and industries

Will they be trustworthy?

- **If the answer must be Yes or No, then the answer is No**
- **Can we develop cyber social learning systems that are trustworthy enough that there is significant benefit associated with their use?**
  - Will these benefits be far greater than the downside costs?
    - Will improvement to quality of life be greater than the costs associated with failures (e.g., loss of life, temporary loss of services, security and privacy violations)
- **Can cyber social learning systems learn to be more trustworthy over time?**

Trust Concerns

- **Reliability**
  - How can we test and validate such systems?
- **Security**
  - How can we develop a CSLS that can thwart most attacks (and ensure a high level of privacy)?
- **Continuous evaluation**
  - How can we monitor the results to determine if they are valid and continue to be valid?

Reliability

- **CSLS will undoubtedly be complex with many different components: control, reasoning, large and growing data sets, human participants**
  - System of systems
  - Numerous examples of failed or poorly designed systems and well functioning systems
- **Numerous testing and verification tools**
  - Strong support for unit testing; infrastructure to support integration testing, etc.
  - Powerful reasoning capabilities for small subsystems
    - But requires considerable investment in resources
      - (E.g. DARPA support to verify the SEL OS kernel**)**

Reliability

- **CSLS will be complex and opaque and thus hard to validate**
  - CS community demanded that the code for electronic voting machines be made publically available and that there be a verifiable voting trail
    - Small, simple systems
    - Can audit the results – know what the results should be!
- **Often will not know if the results are valid**
  - Metamorphic testing tests for “expected” trends
- **CSLS employ ML and other approaches whose accuracy will be hard to determine**
  - What are the properties that should be proven?
- **Humans are unreliable participants and users**
  - Inadvertent errors, malicious actions

Security

- **Results from a CSLS could have enormous economic impact**
  - Findings could influence the choice of medications, medical devices, text books, appliances, fuel combinations, etc.
  - Thus there is the potential for fraud
    - In the design (e.g., Volvo) or through hacks on the system or the data
- **Must demand the use and development of best practices**
  - Development practices: programming languages, coding practices, architectural design, validation
  - Physical security
  - Process safeguards
    - E.g., Limit opportunities for collusion, insider attacks, single points of failure

Continuous Evaluation

- **Results must be continuously questioned**
  - Employ N-version programming
    - Significantly different ML algorithms evaluating the same data; careful analysis of the differences
  - Check for and guard against cultural biases
    - E.g., physician bias impacting the results because of different responses to men versus women or other segments of society
- **CSLS will need to continuously evolve, and be continuously reevaluated**

CSLS raise many hard research questions

- Testing
- Verification
- Security
- Multi-faceted monitoring
- Systematic, validated, and continuous improvement

  - In our enthusiasm for CSLS, Computer Scientists need to be honest about the concerns and be strong advocates for research to address these concerns