Trusted AI and AI Trust
An Opportunity for Synthesis

Munindar P. Singh
singh@ncsu.edu

Department of Computer Science
North Carolina State University

July 2018
Sociotechnical Systems

Current AI research: atomistic, single-agent decision-making and ethical dilemmas
Current social sciences research: Not computational in outlook
Comparison

- Requirements
- Scope
- Aspect
- Autonomy
- Nature
- Fairness
- Research Focus
Comparison

Trusted AI

- Requirements: Of agents to people
- Scope: Trustworthiness
- Aspect: Instrumental: agents are tools
- Autonomy: Intelligence and complexity
- Nature: Transparency
- Fairness: Statistical wrt protected groups
- Research Focus: Individual dilemmas
## Comparison

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Trusted AI</th>
<th>AI Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of agents to people</td>
<td>Trustworthiness</td>
<td>By agents of others</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>Instrumental: agents are tools</td>
<td>Trust</td>
</tr>
<tr>
<td>Intelligence and complexity</td>
<td>Decision making wrt social relationships</td>
<td>Sociocognitive: agents are socially intelligent</td>
</tr>
<tr>
<td>Transparency</td>
<td>Statistical wrt protected groups</td>
<td>Accountability</td>
</tr>
<tr>
<td>Individual dilemmas</td>
<td>Individual wrt vulnerability</td>
<td>Systemic properties</td>
</tr>
<tr>
<td>Research Focus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion
Going back to sociotechnical systems

- Build on sociocognitive modeling
- Incorporate human considerations of interpretability and understanding
- Incorporate reasoning about incentives
- Support composition
  - Systems of systems with . . .
  - Systems that appear as agents
  - Systems that appear as tools
- Toward a theory of ethics