Explainable AI, Transparency, and Human Machine Teaming

Marc Steinberg

Human Factors and Ergonomics Society
Annual Meeting, October 2019

Teaming as a Spectrum of Capabilities

Falconry
-Narrow Task Competence
-Robust to many environments
-Human believes there is bi-directional trust
-Skilled Human can Train (time intensive)
-Can coordinate heterogeneous tasks
-Dominance hierarchies

Working Dog
-Spontaneously Picks up on Many Human Cues
-Limited Task Competencies
-Robust to many environments
-Human Trust within constraints
-Can help human in controlling non-team entities
-Human Can Train (time intensive)
-Dominance related social skills

High-Functioning Mammals
-Understanding of others as animate and directed
-Precursors of or rudimentary theory of mind (controversial)
-Range of social structures & skills
-Emulation learning
-Attention to human cues when raised by humans

Human Children
-Understand intentionality
-Theory of Mind, Perspective, Joint Attention
-Basic Language
-Understanding of events, sequences, narratives
-Imitation learning, teaching others
-Dialogue/Social skills (e.g., turn taking)

High Functioning Adult Human Teams
-Reduced need for explicit communication
-High degree of trust
-Understanding of others actions, affect, intentions & how to influence others mental state
-Understanding subtle social cues
-Joint awareness
-Meta cognition

Cross-Disciplinary Challenges

- Algorithm/Device Focused Communities
- Human Factors/Ergonomics
- Human Computer Interaction (HCI)
- Human Robotic Interaction

There is not yet an integrated community that can fill the spectrum between the expensive, and difficult to apply methods used with high assurance systems, and the more lightweight ones needed in many fast moving applications

Example Approaches

- Does more human-like AI inherently have benefits?
- Constrain, convert, or pair AI methods to ones with some more “natural” human-understandable direct translation
- Adjust AI methods to generate more human-understandable plans and behaviors even if the underlying algorithm is opaque and difficult to understand.
- Meta-cognitive level
  - Direct explanation mechanisms that reason about AI methods
  - Assess and communicate in situ performance and broader proficiency across many similar environments or tasks.
- Decision aids to simulate and predict system behaviors and sensitivities, and explore alternative scenarios and counterfactuals that particularly support skilled users.
- Individual component level vs end-to-end methods for explanation, transparency, etc.