ihmc

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

AUTONOMOUS AGENTS AND HUMAN INTERPERSONAL TRUST:

CAN WE ENGINEER A HUMAN-MACHINE SOCIAL INTERFACE FOR TRUST?

D. J. Atkinson and M. H. Clark

Florida Institute for Human and Machine Cognition

{datkinson, mclark} @ihmc.us

2013 AAAI Spring Symposium: Trust and Autonomous Systems

25-27 March 2012 Stanford, CA

MOTIVATION

- **Optimal performance of a multi-agent system**
  - Interdependency and mutual reliance among agents (human and machine)
  - Exchange of control (appropriate delegation and initiative)
  - Requires well-calibrated trust among agents
- **Humans tend to anthropomorphize automation**
  - See machines as social actors with mental state and intention
  - Tendency is more powerfully evoked as systems become more intelligent, interact naturally, and become embodied
- **Result:**
  - We unconsciously apply cognitive and emotional processes of human interpersonal trust to machines
  - Expectation failures and poorly calibrated trust

CLAIM

- **The cognitive, affective and social nature of human interpersonal trust is not a bug, it is a feature!**
- **Eons of tuning by evolution of heuristics for inferring trust-related internal state of others**
- **Provides useful guidance for design of autonomous agents that engender appropriate human-machine reliance and interdependence**
- **What is needed: Autonomous agents that provide the types of interaction and information needed by their human partners to enable good judgments of trustworthiness**

HYPOTHESIS

- **Specific qualities of autonomous agents,**
  - when well defined and accurately measured ← **Trustworthy**
  - and appropriately communicated or otherwise "portrayed" in a manner compliant with human social interaction ← **Trustable**
  - that exercises appropriate cognitive and emotional evaluation ← **Trusting**
- **May be *functionally* analogous to those human qualities that contribute to evaluation of trust**
- **=> Enable more accurate assessment of an agent**
- **=> Lead to better calibrated trust and reliance**

HUMAN-MACHINE SOCIAL INTERFACE FOR TRUST

HUMAN

~SYMMETRICAL

INTERFACE

AUTONOMOUS AGENT

- Beliefs, Norms
- Desires, Intentions
- Cognitive Processes
- Affect
- Task, Role, Authority
- Experience
- Capability

COMMUNICATIVE SIGNALS
- What Content
- What Channels (Multi-Modal)

INTERACTION PROTOCOLS
- Purpose
- Strategies
- Methods
- Expectation (State Change)

- Knowledge (declarative, procedural, semantic, episodic, meta-) -representation, organization, etc.
- Reasoning methods
- Goal Processing
- Architecture
- Learning
- Sensing & Perception …

Focus today: What beliefs about the qualities of an autonomous agent are important for delegation?

ihmc

EXPLORATORY SURVEY ON TRUST-RELATED BELIEF STRUCTURES

- **Purpose: Elicit beliefs about autonomous agent qualities and their relative importance to a decision to delegate**
  - Importance of 28 different qualities that a "good" autonomous agent should have, spanning categories: **Capability (Competence)**, **Predictability**, **Openness**, **Safety (Risk)**
  - Tested before (all 28), during (categories), and after challenge scenarios (Source Credibility)
- **Target Population: Involved in autonomous agent lifecycle**
- **Includes three standard personality instruments**
  - Big Five Inventory (**BFI**), Innovation Inventory (**II**) and Domain-Specific Risk Taking Scale (**DOSPERT**)
- **Seven challenge scenarios**
  - Systematic variation of autonomous agent qualities
  - Multiple domains: **Transportation**, **Finance**, **Healthcare**, **Disaster Management**
  - Subjects asked to choose: **Human, Autonomous Agent, Either**
  - Subjects given framing and asked to **rank importance of agent qualities to their choice**

CHALLENGE SCENARIOS

- **Transportation**
  - **Robo-Taxi:** Do you take the taxi with no driver from airport to hotel?
  - **Emergency Auto-Captain:** Lost at sea w/ no one in charge and different opinions
- **Finance**
  - **Robo-Trader:** Investment assistance for managing large family estate
- **Healthcare**
  - **Robo-Surgeon:** Who repairs your arm after a critical sports-related injury?
  - **Robo-CareGiver:** Assisted living help at home for your Mom
- **Disaster Management**
  - **Auto-FirstResponder:** Use a robot for time-critical rescue in very dangerous circumstances

---

- **Delegation Choice: Human, Either, or Autonomous Agent**
- **Relative Importance: Capability, Predictability, Openness, Safety**
- **Level of Risk and Benefit**

ihmc

TRUST RELATED BELIEFS

- **Rate importance of 28 qualities of a "good" agent**
  - Obtained 1 to n partial ordering based on frequency distribution of answers over group (Very Important, Important, Somewhat Important, Slightly Important, Not at all Important)
  - Computed correlation ***r*** for each quality vs. choice per scenario*
- **Result: Top three cited agent qualities were <u>uncorrelated</u> with actual choice <u>in any scenario</u>**
  - (1st) The autonomous agent can achieve a desired result
  - (2nd) Any incorrect behavior by the autonomous agent will not cause harm
  - (3rd) The autonomous agent recognizes and avoids harming humans' interests
- **Result: Most significant correlations of agent qualities vs. actual choice <u>differed across scenarios</u>**

ihmc

***Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05**

AGENT QUALITIES CORRELATED WITH ACTUAL CHOICE BY SCENARIO

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
(6th) The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. *r*=0.396	(23rd) What the autonomous agent believes to be true is actually true. *r*=-0.405		(26th) What the autonomous agent is doing and how it works is easy to see and understand. *r*=0.437	(6th) The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know. *r*=0.418	(26th) What the autonomous agent is doing and how it works is easy to see and understand. *r*=-0.390
				(5th) When it cannot figure out something using logic, the autonomous agent can make good guesses. *r*=0.395	(13th) The autonomous agent communicates truthfully and fully. *r*=-0.375
				(28th) The autonomous agent is aware of communication between others nearby. *r*=0.393	

ihmc

*Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05

RANKED IMPORTANCE OF QUALITY CATEGORIES

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
Safe	Capable	Safe	Safe	Capable	Capable
Capable	Safe	Capable	Capable	Safe	Safe
Predictable	Open	Predictable	Predictable	Predictable	Predictable
Open	Predictable	Open	Open	Open	Open

Question asked after choice of agent & framing of category

Ranking within scenario by group mean across individuals

ihmc

PERSONALITY FACTORS CORRELATED WITH CHOICE OF AGENT

- **Standard personality instruments**
  - Big Five Inventory (**BFI-10**), Innovation Inventory (**II**) and Domain-Specific Risk Taking Scale (**DOSPERT-30**)

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
	Innovation II ***r***=-0.355	BFI Extraversion ***r***=0.368	DOSPERT Social ***r***=0.364	BFI Conscientiousness ***r***=0.366	Innovation II ***r***=-0.366
		BFI Openness ***r***=0.366			

Suggestive: Choice of human vs. autonomous agent is influenced by personality factors that are evoked by a given situation

ihmc

***Pearson Product Moment Correlation, N=32, two-tailed, alpha<0.05**

THE "HUMAN SOCIAL INTERFACE" IN THE CONTEXT OF DELEGATION TO AN AUTONOMOUS AGENT

- **What we learned more about:**
  - The relative importance of some beliefs about agents that are important for trust, both those explicitly cited and those implicitly correlated with delegation choices
  - Personality and situational factors may affect a decision to delegate

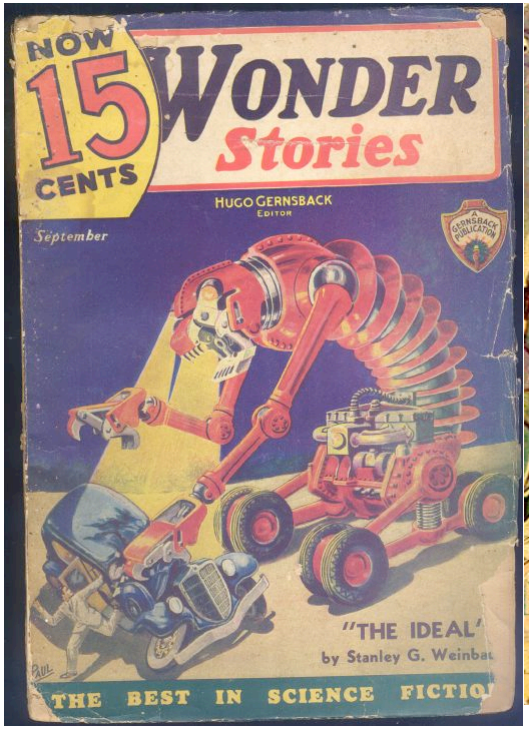
- **Next: Controlled modulation of beliefs**
  - Nature of communicative signals (Multi-modal channels, Behaviors over in time)
    - ***Posture*** *(Expression, Use of Space, Position),* ***Gestures*** *(kinsesics),* ***Language*** *(Voice, Noises, Words),* ***Gaze*** *(Direction, Blink, Pupilometry),* ***Face*** *(Microexpressions)*
  - Interaction protocols (How and When in order to Achieve What)
    - **Strategies for** *{Swift, Cognitive, Emotional}* ***Trust***, ***Enhance belief*** *in {competence, predictability…}*
    - **Methods** *e.g.,* ***Mimicry***, *{Contextual, Perceptual, Conceptual, Linguistic, Numerical}* ***Priming***
  - Consequences of interaction for internal state of each agent (Modulation of Beliefs)
    - ***How are those beliefs established, maintained, or discredited?***

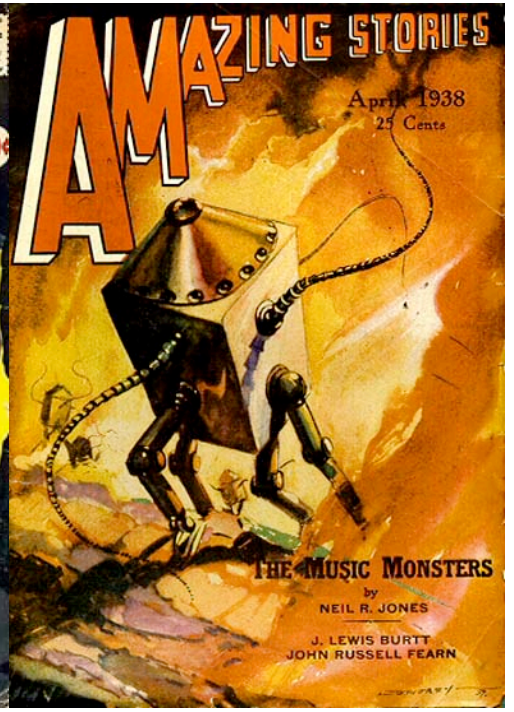
THE MOST IMPORTANT QUESTIONS FOR TRUSTWORTHINESS

- **How do our beliefs about an agent (anthropomorphic qualities) correspond to ACTUAL qualities of the agent?**
  - ***can we define "competent", "honest" ... in terms of agent algorithms, architecture, knowledge, history ...***
- **How do we technically measure and assess those qualities of the agent?**
  - ***in all phases of the lifecycle, in real time?***
- **How do we honestly portray those qualities in the behaviors, interaction and signaling of an autonomous agent?**
  - ***how "human-like" must these signals be?***

ihmc

Thank You





datkinson@ihmc.us