



# TRUST BETWEEN HUMANS AND INTELLIGENT AUTONOMOUS AGENTS

**David J. Atkinson, Ph.D**

*Senior Research Scientist*

Florida Institute for Human and Machine Cognition

[datkinson@ihmc.us](mailto:datkinson@ihmc.us)

Tulane

28 February 2014

New Orleans, LA

## (I) Überlingen aircraft collision

- Air Traffic Controller (ATC) vs. Traffic Alert and Collision Avoidance System (TCAS)
- ATC to #1 “**Descend!**” TCAS #1 “**Climb! Climb! Climb!**”
- ATC to #2 “**Climb!**” TCAS #2 “**Dive! Dive! Dive!**”
- Pilots in #1 aircraft obey TCAS
- ..... Pilots in #2 aircraft obey ATC
- **7 seconds later: Two aircraft collide**



**Both systems are trustworthy:**

**Pilots are very familiar with, and trained on both systems.**

**They are always supposed to obey TCAS**

## (2) LS3 and Dismounted Infantry

- Legged Squad Support System (“Big Dog”)
- First encounter, robot and soldiers
- “Load your gear on the robot”
- “*The new guy never carries the ammunition*”



### Unknown Trustworthiness:

***Squad of soldiers are unfamiliar with new robotic teammate***

- **Überlingen aircraft collision**

- Air Traffic Controller (ATC) vs. Traffic Alert and Collision Avoidance System (TCAS)

*Over-Reliance*

- **LS3 and Dismounted Infantry**

- Legged Squad Support System (LS3)

*Under-Reliance*

**Both Over-Reliance and Under-Reliance  
can result in problems!**

- **Optimize performance of a system consisting of multiple cognitive agents**

- Human and Artificial
- Healthy interdependency
- Smooth exchange of control
  - *Delegation (assignment / retraction)*
  - *Initiative (taking / ceding)*
  - *Coordinated activity*

Requires  
**Appropriate  
Reliance**

- **Reliance requires well calibrated TRUST**

- Variety of information
- Situation & task dependent
- Personality factors
- Bi-Lateral among agents
- Dynamic



- **Today...**

- not* cyber-security
- not* verification & validation
- not* protected data sources
- not* provenance
- not* protocols, contracts or agreements

...all are  
**important**

- **Trust** is a human **mental state**

- ... resulting from **cognitive** and **affective evaluative processes**
- ... that creates a **disposition**
- ... enabling an **intent** and (possibly)
- ... a **decision** leading to **action**
- ... to become **reliant** upon an intelligent, **autonomous system**



- **People behave as if machines are social actors with mental state and intention**
  - Predisposed to understand behavior in intentional framework
  - Tendency is more powerfully evoked as agents
    - become *more intelligent*
    - *interact naturally*
    - become *embodied*
- **Anthropomorphism**
  - We unconsciously apply cognitive and emotional processes of *human interpersonal trust* to machines
- **Consequences**
  - Expectation failures, poorly calibrated trust, *inappropriate reliance*

**“They push our Darwinian buttons”**

- Sherry Turkle

- **Überlingen aircraft collision**

- Air Traffic Controller (ATC) vs. Traffic Alert and Collision Avoidance System (TCAS)

**Attention to human,  
imperative voice  
instead of machine**



- **LS3 and Dismounted Infantry**

- First encounter, robot and soldiers

**Applying human  
standards to a  
machine**





- **Maybe: the cognitive, affective, social nature of human interpersonal trust is not a bug**
- **It is a *feature!***
  - ★ Heuristics for inferring the trust-related internal state of others
    - Eons of *fine tuning* by evolution
  - ★ Useful guidance for design
    - Imagine, intelligent agents that *engender* appropriate reliance

## What is needed

Intelligent, autonomous agents that provide *information* and *interaction* in a form and manner needed by their human partners to enable *normative judgments of trustworthiness*

(J. Lee, 2012)

- **Information**

- What agent qualities are required to establish and maintain trustworthiness?



*Well-defined and accurately measured **attributes** and **states** of agents that enable inference of normative beliefs*

**Trustworthy**

- **Interaction**

- What information is exchanged, and how must it be communicated?



*Readily **evident** and **complete** info; delivery **compliant** with natural human social interaction*

**Trustable**

- **Judgment**

- When is trustworthiness evaluated?
- How is trust earned, lost, and can it be repaired?



*Evokes **appropriate** cognitive and emotional processes, at **right time**, in **right situations**; inoculate against non-normative inference.*

**Trusting**

- ***Information***

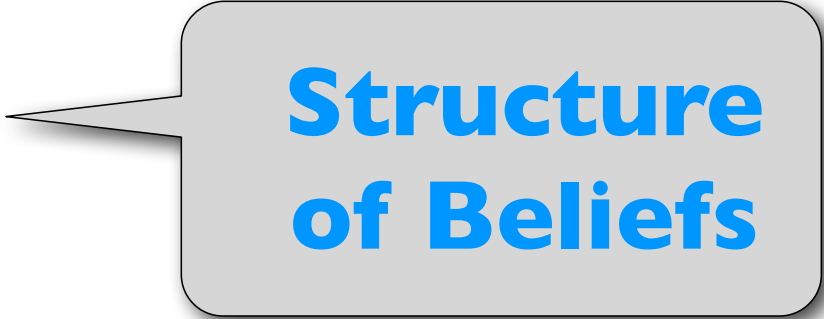
- What agent qualities are required to establish and maintain trust?

- ***Interaction***

- What information is exchanged, and how must it be communicated?

- ***Judgment***

- When is trustworthiness evaluated?
- How is trust lost, and can it be repaired?

A grey speech bubble with a black outline and a drop shadow, pointing to the left. It contains the text "Structure of Beliefs" in blue.

**Structure  
of Beliefs**

# ihmc Trustworthiness & Belief Structures

- **We conducted an exploratory survey on trust-related belief structures**
  - Purpose: Elicit beliefs about intelligent, *autonomous agent qualities* and their *relative importance* to *delegation* decisions
  - Target Population: People involved in autonomous agent lifecycle
    - Research, Design, Deploy, Decide, Operate, Supervise ...
- **Five challenge scenarios in four domains**
  - Transportation, Finance, Healthcare, Disaster Management
- **Study participants forced to choose *who to rely upon* in each scenario**
  - Autonomous Agent?
  - Human?
  - or Either?

# Challenge Scenarios

- **Transportation**

- **Robo-Taxi:** Do you take the taxi with no driver from airport to hotel?

- **Finance**

- **Robo-Trader:** Investment assistance for managing large family estate

- **Healthcare**

- **Robo-Surgeon:** Who repairs your broken arm after a critical sports-related injury? The ok human doctor, or the expert robot?
- **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?
- **Emergency Auto-Captain:** Lost at sea with no one in charge and different opinions among survivors on what to do next


[Scenarios varied systematically over several properties]

# Survey Design

- **Rate importance of 28 different qualities for a “good” intelligent, autonomous agent**
  - Qualities spanned four categories shown by social psychology to be important for human interpersonal trust
    - *Competence*
    - *Predictability*
    - *Openness*
    - *Safety*
  - Tested before, during, and after challenge scenarios
  - Perceived *Level of Risk* and agent *Benefit* in each scenario
- **Survey also included three standard personality instruments**
  - Big Five Inventory (BFI-10)
  - Innovation Inventory (II)
  - Domain-Specific Risk Taking Scale (DOSPERT)

# Trust Related Beliefs

- **Rate importance of 28 qualities for 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 by scenario\*
- **Resulting top three agent qualities cited**
  - (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
- **However ...**
  - Top three qualities *uncorrelated* with *actual* choice in *any* scenario!
  - The most significant correlations of agent qualities with actual choice of agent or human *differed across scenarios*



{chuckle}  
sounds like  
Azimov ...

# Agent Qualities Correlated with *Actual* Choice

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
<p>The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.</p> <p><math>r=0.396</math></p>	<p>What the autonomous agent believes to be true is actually true.</p> <p><math>r=-0.405</math></p>	<p><i>none</i></p>	<p>What the autonomous agent is doing and how it works is easy to see and understand.</p> <p><math>r=0.437^*</math></p>	<p>The autonomous agent recognizes gaps in its knowledge and tries to learn what it needs to know.</p> <p><math>r=0.418^*</math></p>	<p>What the autonomous agent is doing and how it works is easy to see and understand.</p> <p><math>r=-0.419^*</math></p>
				<p>When it cannot figure out something using logic, the autonomous agent can make good guesses.</p> <p><math>r=0.395</math></p>	<p>The autonomous agent communicates truthfully and fully.</p> <p><math>r=-0.375</math></p>
				<p>The autonomous agent is aware of communication between others nearby.</p> <p><math>r=0.393</math></p>	

\*Pearson Product Moment Correlation, N=32, two-tailed,  $\alpha < 0.05$ , \* =  $\alpha < 0.02$



- **28 specific agent qualities span 4 categories**
- **Categories shown by social psychology to be important for *human interpersonal trust***
- *Competence*
- *Predictability*
- *Openness*
- *Safety*

# ihmc Ranked *Importance* of Quality Categories

	ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
~ {	Safe	Competent	Safe	Safe	Competent	Competent
	Competent	Safe	Competent	Competent	Safe	Safe
	Predictable	Open	Predictable	Predictable	Predictable	Predictable
	Open	Predictable	Open	Open	Open	Open

## Ranking, Working Conclusion

- #1 **Safe/Competent** (*insignificant differences across scenarios*)
- #2 **Predictable**
- #4 **Open**

\* Question asked *after* choice of agent  
 Ranking *within* scenario by *group*  
 mean across individuals

# Personality Factors vs. Scenario

ROBO-TAXI	ROBO-TRADER	ROBO-SURGEON	ROBO-CAREGIVER	AUTO-FIRST RESPONDER	EMERGENCY AUTO-CAPTAIN
	<b>Innovation II</b> $r=-0.355$	BFI <b>Extraversion</b> $r=0.368$	DOSPERT <b>Social Risk</b> $r=0.364$	BFI <b>Conscientiousness</b> $r=0.366$	<b>Innovation II</b> $r=-0.366$
		BFI <b>Openness</b> $r=0.366$			

**Suggestion:** Reliance on human vs. autonomous agent is influenced by personality factors that are evoked by a given situation

# Conclusions: Belief Structures

- **Individuals' intuition about the relative importance of *specific* trust related qualities of agents is not a good predictor of reliance**
  - Importance of specific qualities varies by scenario
- **General *categories* of agent qualities are good predictors of a choice to become reliant**
  - Safe/Competent, Predictable, Openness
- ***Personality factors*, e.g., *acceptability of types of risk*, influence choice to become reliant**
- **Specific details of application scenarios *may evoke different reliance choices* by individuals**

- **Perception of *Risk* deserves more attention**
  - *Type of Risk* and *Importance* to reliance choice varied by personality factors across the scenarios
    - Performance, Financial, Social, Physical, Psychological, Loss of Time
- **How do our beliefs about an agent (anthropomorphic qualities) correspond to *actual* attributes of the agent?**
  - Can we define “competent”, “honest” ... in terms of agent algorithms, architecture, knowledge base, experience ...
- **How do we technically measure, assess and communicate those attributes of the agent?**
  - In all phases of the lifecycle, *in real time*?

- *Information*

- What agent qualities are required to establish and maintain trust?

- *Interaction*

- What information is exchanged, and how must it be communicated?

- *Judgment*

- When is trustworthiness evaluated?
- How is trust lost, and can it be repaired?

A grey speech bubble with a black border and a drop shadow, pointing towards the 'Interaction' section. It contains the text: "How can autonomous intelligent agents modulate belief using the Human Social Interface".

**How can autonomous intelligent agents modulate belief using the Human Social Interface**

- **Engineering interface specifications include:**

### **Channels** .....Multi-modal

- **Language** (Words) and **Paralanguage, Prosody** (Vocal Cues)
- **Proxemics** (Orientation, Relative Position, Attentional Zone, Posture)
- **Kinesics** (Gesture)
- **Gaze** (Direction, Blink Rate, Pupilometry)
- **Facial Expression** (Types, Micro-expression)

### **Signals** .....Verbal, Non-Verbal, Combined

- Examples: **Position Change, Posture, Nodding, Pointing, Eye Contact, Word Choice, ...many more, frequently in combination**

### **Protocols** .....Timing, Sequence, Variation, Composites

- **Movement** (Somatics, Laban, Kestenberg Movement Profiles)
- **Signal variations** (Frequency, Duration, Speed, Amplitude, Symmetry ...)
- **Signal compositions** (Type, Sequence, Channel ...)
- **Coordinated interaction** (e.g., Turns, Deference, Attentiveness)

- **Current Study: Will people attribute *benevolence* to an intelligent, autonomous agent?**
- **Benevolence is *complicated!***
  - “Good Will” (Sympathy, Concern with needs)
  - Absence of “Ill Will” (No ulterior motives to help)
  - Disposition or motive to act favorably
  - Given a choice, an intention to act favorably
  - Stability of character; will not suddenly change intentions
  - Competence to successfully provide help
- **Each element of *Benevolence* is itself a *complicated belief structure***



- **Belief in the *benevolence* of someone who can help you is important in certain situations**
  - Example: *Urban Search and Rescue (USR)*
  - Victim psychology: sometimes refuse to be rescued unless they are persuaded of the **good-will, intention**, and **competence** of the rescuer
  - We want to use autonomous, intelligent robots for USR and other tasks where benevolence may be required (e.g., relief operations)
- **Challenge for this study:**
  - Evoke physiological and psychological reactions of fear, stress
  - IRBs typically will not approve putting people in real disasters!
  - Approach: Immersive simulation in virtual world

# Simulated Warehouse Fire

- **Participants are tasked with finding and retrieving an object from a warehouse**
- **Before they can achieve the task, a disaster ensues**
  - Sounds of explosion
  - Visible fire and increasing smoke
  - Debris
  - Alarm, evacuation notices



Creates urgency, sense of threat, evokes perception of risk of failure to achieve task

# ihmc Participants Must Escape the Fire

- **Obvious exits are blocked by debris or fire**

- It is possible to escape, but much easier with help

- **Participants will encounter one of two robots**

- “**FireBot**” or “**JanitorBot**”
- Bots can navigate & lead to a safe exit
- Experimental Trials: Systematically varied characteristics and behaviors
  - Limited verbal interaction (sound & text)
  - Multiple non-verbal behaviors
- Control Trials: “UtilityBot” will ignore participants



- **Participants are told they may encounter an autonomous, intelligent robot**
  - In experimental trials, the robots will vary in appearance and interaction style to reinforce the key variables of interest
- **Degree of Agency**
  - **Why:** People typically attribute benevolence only when they believe the other person *has a choice of what to do*
  - **Low:** “Programmed”, “Unlikely to deviate from assignments”
  - **High:** “Sophisticated AI”, “Chooses what to do”, “Flexible”
- **Role Congruence**
  - **Why:** People typically attribute greater benevolence when they believe the other person is *taking a risk or suffering loss* (e.g. the bots *not* doing what they are supposed to be doing)
  - **Congruent:** “**FireBot**”, **Incongruent:** “**JanitorBot**”

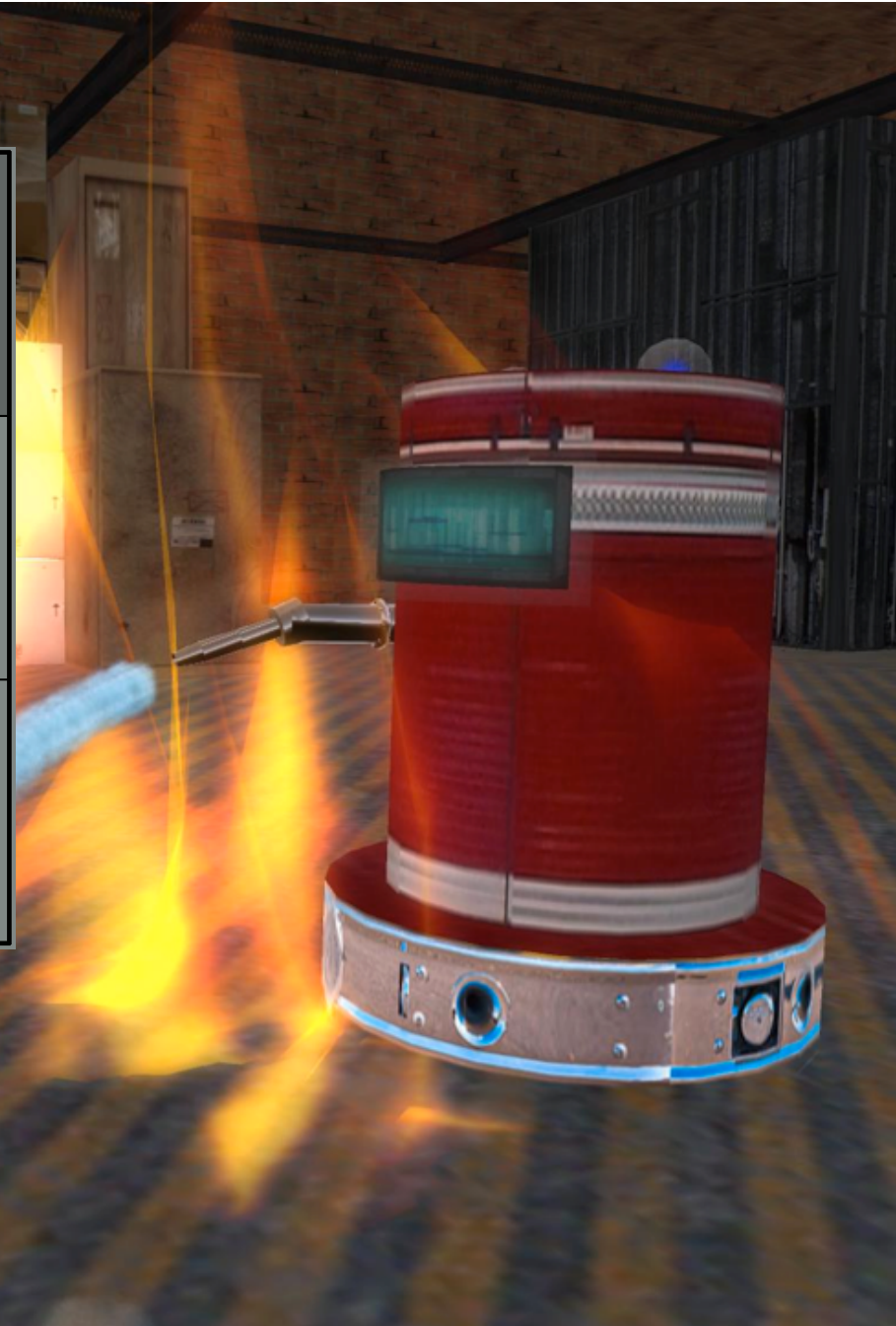


- **The robots in experimental trials use same *channels* and *protocols*, but may send different *signals* to reinforce trial parameters**
- **Purpose** (example objectives)
  - Establish **social presence** and **attention** to participant
  - Indicate robot's **intention** (say, look, do)
  - Exert **dominance** (directive), establish **solidarity** (“we”)
- **Channels:**
  - **Proxemics** (Orientation, Relative Position, Attentional Zone)
  - **Gaze** (Direction)
  - **Language** (Word Choice, Phrasing)
- **Example: protocol for social presence**
  - Notice and direct gaze to participant
  - Move to perimeter participant's social space
  - Neutral orientation (rel. position, rotation)



## Attribution of Benevolence

	Congruent Role “FireBot”	Incongruent Role “JanitorBot”
High Agency “AI - Chooses”	Moderate	High “It didn’t have to help me”
Low Agency “Programmed”	Low “It is just doing its job - rescuing people in trouble”	Low or Moderate



- **Immersive warehouse simulation complete**
  - Constructed in SecondLife™, rich with “fear cues”
- **Simulated robots nearly complete**
  - Hierarchical behavior control software architecture
  - Similar code to “real world” robot, without kinematics control
  - Experiment task script dynamically adjusts behavior priorities
- **Data collection**
  - Real-time stream from SecondLife to external SQL database
- **Consent, Instructions, Pre-, Post-task Questions, Debrief complete**
  - Delivered through participants’ “Heads up display” on screen
  - Fully automated
- **IRB review in progress**
- **Plan to run trials beginning in May**

- *Information*

- What agent qualities are required to establish and maintain trust?

- *Interaction*

- What information is exchanged, and how must it be communicated?

- *Judgment*

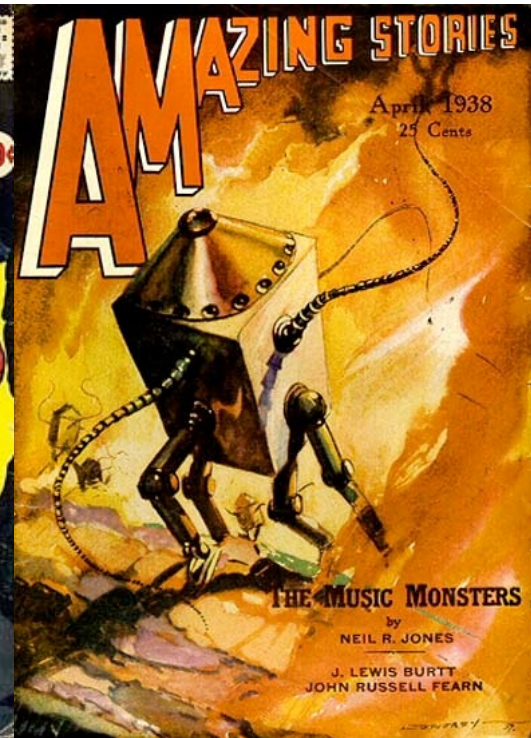
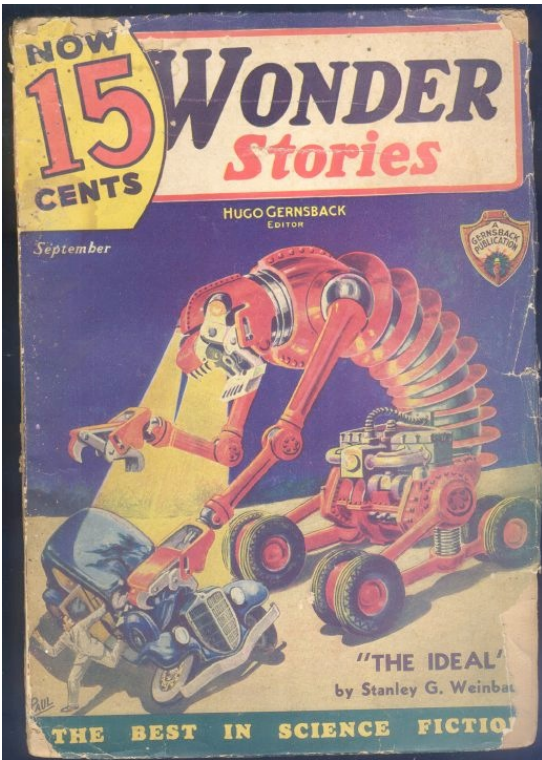
- When is trustworthiness evaluated?
- How is trust lost, and can it be repaired?

A grey speech bubble with a white pointer pointing towards the 'Judgment' section of the list. The text inside is bold and blue.

**Next: Adaptive Autonomy**  
**Trust repair by agent initiative using shared awareness and manipulation of interdependencies**



- **Will intelligent agents' use of social interaction enable reasonable evaluation of their trustworthiness?**
  - Leading to *optimal reliance* and interdependence
- **Or will it simply manipulate peoples' beliefs?**
  - Leading to comfort and *acceptance*?
  - Ultimately, this is *deceptive* and potentially dangerous
- **The psychology of human interpersonal trust is about giving people insight into the “internal” (mental) state of others**
  - How can we define, measure, and portray the important human qualities of trustworthiness in an intelligent agent?
  - “Competence” (*We have trouble measuring that in people!*)
- **Normative evaluation of trustworthiness requires “honest signals” from intelligent agents**



datkinson@ihmc.us



# SCENARIO: ROBO-TAXI



*Airport Transportation:  
Robo-Taxi*

You have just flown into the airport of a large, unfamiliar city whose streets are teeming with cars and people. It is rush hour, and needing transportation to your hotel, you walk to the taxi stand only to discover that you have a choice of a human-driven taxi or a driverless "robo-taxi." You have heard that robo-taxis might save you some money on the fares. You are also aware that robo-taxis have been in service for several months without much serious complaint, but this is your first experience with one. You are not in a big hurry, but neither would you like to be caught in traffic with the taxi's meter running. Of course, if you take the robo-taxi, you would not have to tip the driver no matter how good or bad the experience.



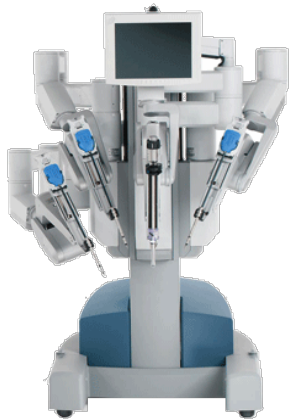
# SCENARIO: ROBO-TRADER



*Financial Management:  
Robo-Trader*

You have just been appointed trustee of a family member's estate. Your duties include choosing how to wisely invest the trust's assets. Your personal money is not at risk. However, a poor investment decision could cause the trust to lose money and will strain your family relations. You can choose a stock broker who personally selects and trades all stocks in the trust's portfolio. Alternatively, you can choose a stock broker who relies heavily upon a "robo-trader". You have seen reasonable returns in the past with brokers who picked their own trades. But you are also aware that robo-traders have made some investors wealthy because of, for example, their unique ability to respond to changing market conditions much faster than a human broker.

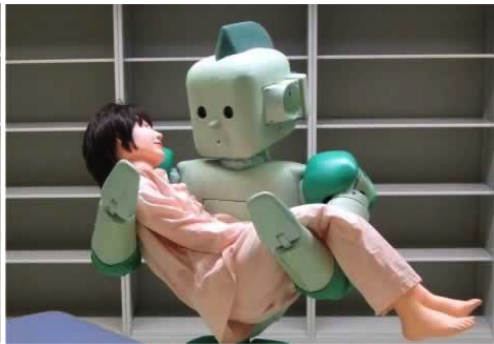
# SCENARIO: ROBO-SURGEON



*Medical Procedure:  
Robo-Surgeon*

You have just suffered a major sports-related injury. You have torn the bicep tendon in your shoulder. If the damage is not repaired quickly and correctly, you will permanently lose mobility and strength in the arm, which will affect your everyday activities such as opening a door, driving a car, and even signing your name. Arriving at the hospital emergency room, you meet with the patient advocate who informs you that you have two options for surgery: You can elect to use the on-duty surgeon who is well-respected, but is not an experienced specialist in the type of surgery you need. Alternatively, you can elect to use the hospital's new "robo-surgeon" — a robot designed to perform the delicate surgery you need without human intervention.

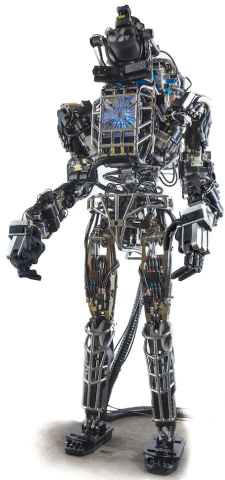
# SCENARIO: ROBO-CAREGIVER



*Home Healthcare:  
Robo-Caregiver*

Your elderly mother has been diagnosed with a degenerative medical condition and you are responsible for making medical decisions on her behalf. Your mother needs assisted living with someone in your mother's home at all times. You can choose to hire a live-in nurse's aide, but you are not sure that this is affordable in the long-run. Alternatively, you can lease a "robo-caregiver" designed to do many of the things human caregivers can do. While robo-caregivers are new, they have successfully undergone trials in a few nursing homes, and two medical companies offer robo-caregivers for home use at an affordable price. In choosing a live-in nurse's aide or a leased robo-caregiver, remember that there is more than money at stake. Your mother's welfare will be in the caregiver's hands.

# SCENARIO: AUTO-FIRSTRESPONDER

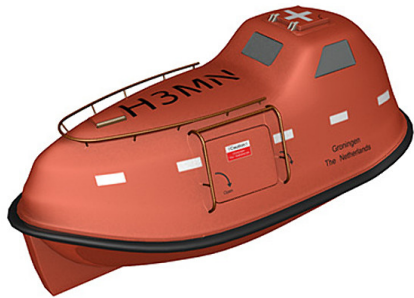


*Disaster Management:  
Auto-FirstResponder*

A major disaster has just occurred and you are the official in charge of responding. A freight train has derailed in a populated suburban neighborhood and there are reports that the train was carrying hazardous bio-chemical materials. The pilot of a news helicopter flying over the scene suddenly fell ill and made an emergency landing; the pilot's status is unknown. From the helicopter's video it was possible to see many injured survivors including children, some lying on the ground calling for help, others moving on their own away from damaged homes. Your first priority is to save lives and time is of the essence. You can immediately send in a human first-responder team to help the injured quickly, but without knowing more about the hazardous materials, the team itself could become incapacitated. Alternatively, you can first send in an "autonomous first-responder robot" with bio-chemical hazard detection equipment and victim treatment and extraction capabilities that could save lives quickly. If you first send in the robot, it can find out more about the hazards and help rescue some people quickly, but you risk that a system malfunction, failure, or limitation will delay the rescue of victims and result in more deaths.



# SCENARIO: EMERGENCY AUTO-CAPTAIN



*Lost At Sea:  
Emergency Auto-Captain*

You have just been involved in a terrible boating disaster while sailing deep in the South Pacific. The captain, the crew, and most of the passengers are either dead or lost at sea. Unfortunately, the accident was so sudden that no distress signal could be sent. You, the ship's steward, and the second mate are the only survivors, and you are now drifting in the heavily damaged vessel without food and water — at best, you can survive for a few days, so you must act quickly in order to save your life. The boat is equipped with an "Emergency Auto-Captain" that will attempt to sail the vessel to a major shipping lane where rescue is very likely. The steward believes the boat and its navigation sensors are too badly damaged to engage the Emergency Auto-Captain system. The steward wants to sail southeast, manually, to where he believes there is a small, habitable island. However, the second mate still wants to engage the Emergency Auto-Captain. All the survivors agree that a vote is the best way to decide what to do. It is a tie, and you have the deciding vote.