Welcome
2017 Faculty Summit Attendees

Faculty Summit 2017
microsoftfacultysummit.com

Microsoft Research
Microsoft.com/research

Facebook
@microsoftresearch

Twitter
@MSFTResearch
#FacSumm
#EdgeofAI
Challenges of Human-Aware AI Systems

Subbarao Kambhampati
BREAKING NEWS

AI HELPS OLD LADY CROSS STREET!

16:43  AI PLAYS WITH KIDS, COOKS FOOD, AND HANGS AROUND SANS DRAMA
Al’s Curious Ambivalence to humans...

• Our systems seem happiest
  • either far away from humans
  • or in an adversarial stance with humans

You want to help humanity, it is the people that you just can’t stand
If we force them to cohabit, they either seem to run us down or commit hara-kiri.
We need to stop freaking out Elon!!

OR NONE OF YOU WILL GET YOUR PREORDERED TESLAS!
Why intentionally design a dystopian future and spend time being paranoid about it?

AAAI-18 Special Track on Human-AI Collaboration!
JASON Briefing on “The Path to General AI goes through Human-Aware AI”; June 2016

8.2 Recommendations

JASON offers the following recommendations to DoD senior leadership:

1. DoD should both track (via a knowledgeable cadre) and invest portfolio) the most dynamic and rapidly advancing areas of AI means limited to DL.

2. DoD should support the development of a discipline of AI eng progress of the field through Shaw’s “craft” and (empirical) “c particular focus should be advancing the “illities” in support of AI.

3. DoD’s portfolio in AGI should be modest and recognize it as an advancing area of AI. The field of human augmentation via AI and deserves significant DoD support.

4. DoD should support the curation and labeling for research, of large data sets. Wherever possible, operational data should be use in support of AI for DoD-unique missions.

5. DoD should create and provide centralized resources for its int researchers (MOSIS-like), including labeled data sets and access training platforms.

6. DoD should survey the mission space of embedded devices for applications of AI, and should consider investing in special use AI inference in embedded devices for DoD missions if identified.

Seeking new algorithms for human-aware AI

Over the years, AI algorithms have become able to solve problems of increasing complexity. However, there is a gap between the capabilities of these algorithms and the usability of these systems by humans. Human-aware intelligent systems are needed that can interact intuitively with users and enable seamless machine-human collaborations. Intuitive interactions include shallow interactions, such as when a user discards an option recommended by the system; model-based approaches that take into account the users’ past actions; or even deep models of user intent that are based upon accurate human cognitive models. Interruption models must be developed that allow an intelligent system to interrupt the human only when necessary and appropriate. Intelligent systems should also have the ability to augment human cognition, knowing which information to retrieve when the user needs it, even when they have not prompted the system explicitly for that information. Future intelligent systems must be able to account for human social norms and act accordingly. Intelligent systems can more effectively work with humans if they possess some degree of emotional intelligence, so that they can recognize their users’ emotions and respond appropriately. An additional research goal is to go beyond interactions of one human and one machine, toward a “systems-of-systems”, that is, teams composed of multiple machines interacting with multiple humans.

Human-AI system interactions have a wide range of objectives. AI systems need the ability to represent a multitude of goals, actions that they can take to reach those goals, constraints on those actions, and other factors, as well as easily adapt to modifications in the goals. In addition, humans and AI systems must share common goals and have a mutual understanding of them and relevant aspects of their current states. Further investigation is needed to generalize these facets of human-AI systems to develop systems that require less human engineering.

NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN
But isn’t this cheating?

Doesn’t putting human in the loop dilute the AI problem?  
Won’t it be cheating?  
Like the original Mechanical Turk...
  --First “human-in-the-loop AI” 😊

NO!
Expands reach and scope of AI enterprise  
Reduces some of the off-the-top worries about AI  
Brings up novel research challenges
Architecture of an Intelligent Agent

- Sensors
  - State
  - What the world is now?
  - How the world evolves
  - What my action does to the world
  - Goals
  - What should I do next?
  - What happens when I do an action?

Environment

Robot

Microsoft Research
Faculty Summit 2017
The Edge of AI
Architecture of an Intelligent Agent teaming with a human

HMM = Human Mental Model

Robot

Human

Environment

Actuators

Goals

Human Goals

State

HMM

What the world is now?

What the HMM is right now?

How the world evolves

How the HMM evolves

What happens when I do an action?

What happens when the human does an action?

What should WE do next?

What my action does to the world

What my action does to the HMM

Sensors
Human-in-the-Loop Planning
Intention Recognition with Emotive
Intention Projection with Hololens
Use Case Scenarios: One Robot & One Human

Prediction: H is about to break a door open when R notices H's intention and predicts that breaking the door open will cause a board will fall on H. R thus moves to catch the board preventively.

Capability models: R notices that a heavy object blocks the entrance to a hallway that H wants to explore. Based on its capability model of H (i.e., what H can and cannot lift) and H's goal, R decides to interrupt its current activity and move the block out of the way.

Mitigate Risk: H needs to search a building for wounded people but is uncertain about the structural integrity and worried that parts might collapse. After communicating this to the robot, the robot proposes a plan that has the robot go in first to assess the risk better and then to split the search in ways that minimize human risk.

Anticipation: R is tasked to wait outside a building and watch out for enemies while H is performing a search inside. After more time has elapsed than it would take to perform the search, the R decides to go inside and find H to help H in case H has encountered any problems.

Normative behavior: R has an order to deliver medical supplies to H, but notices on the way a wounded victim that needs medical attention. R decides that caring for the victim is most important that delivering the supplies and notifies H of the delay.

Coordination based on Mental Model Inference: R learns that H needs to get a medical kit to be able to triage a recently discovered victim. R knows that H is aware that a medical kit is located in a particular room, but infers that H is unaware that R has a medical kit that H could use. While R cannot directly deliver its medical kit or wait for H due to other commitments, it can place its medical kit along the hallway where it expects H to go in order to get the first medical kit, thus relying on H's ability to notice the available medical kit and pick it up instead of the other remotely located kit.

[With Matthias Scheutz]
Use Case Scenarios: Multiple Robots & Humans

**Belief revision:** H1 and H2 are working in two different areas each assuming that the other will take care of a third area. R detects that discrepancies in its mental models of both humans in conjunction with its observations and decides to work on the third area (alternatively, R informs both H1 and H2 about the discrepancy).

**Mental state inference:** R notices that H1 cannot see H2 is approaching with equipment that H1 needs. R further observes that H1 is about to talk to another person and infers that this might be to order the urgently needed equipment from another person. Hence, R contacts H1 directly and informs H1 of H2's arrival.

**Workload:** R knows that H1 has currently high workload (e.g., from running simulations of H1's current activities based on R's model of H1's performance obtained from prior training) and thus does not interrupt H1 with a request from H2 that can wait, but communicates to H2 that it will take care of the request later.

**Social regulation:** R notices an escalation in the interaction between H1 and H2 about how to best proceed, where H1 and H2 each propose different plans. To mitigate, R proposes a compromise plan that contains elements of both H1's and H2's proposals (a "social" solution).

**Activity coordination through shared mental models:**
Two humans H1 and H2 each work with a robotic teammate R1 and R2 in a first responder scenario in the "hot zone" of a natural disaster. H1 and R1 work on the fair side of the designated area, while DHT2 and R2 begin working in an area closer to the boundary. When R1 arrives with H1 at the designed area, R1 notices that not all the necessary equipment is available and communicates with R2 about the availability of the missing items. R2 quickly predicts equipment needs and anticipates that those items are not needed for a while. After quickly getting the OK from H2 to lend the equipment to R1, R2 drives off to meet R1 half-way, exchanges the equipment, and R1 returns to H1 in time to be able to continue triaging the victims with the missing equipment (which H1 did not even notice). Once the equipment is no longer needed, R1 meets up with R2 again, returning the equipment in time for H2 to have it available.
AI Challenges in Human-Robot Cognitive Teaming

Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, Yu Zhang

(Submitted on 15 Jul 2017)

Among the many anticipated roles for robots in future is that of being a human teammate. Aside from all the technological hurdles that have to be overcome on the hardware and control sides to make robots fit for work with humans, the added complication here is that humans have many conscious and subconscious expectations of their teammates -- indeed, teaming is mostly a cognitive rather than physical coordination activity. This focus on cognitive coordination, however, introduces new challenges for the robotics community that require fundamental changes to the traditional view of autonomous agents.

In this paper, we provide an analysis of the differences between traditional autonomous robots and robots that team with humans, identifying the necessary teaming capabilities that are largely missing from current robotic systems. We then focus on the important challenges that are unique and of particular importance to human–robot teaming, especially from the point of view of the deliberative process of the autonomous agent, and sketch potential ways to address them.

Subjects: Artificial Intelligence (cs.AI)

Cite as: arXiv:1707.04775 [cs.AI]

(or arXiv:1707.04775v1 [cs.AI] for this version)
Interaction Requires Modeling the Human

**Explicability:** Aim to get $\pi^R$ closer to $\pi^H$ (by getting $M^R$ closer to $M^H$)

**Explanation:**
Tell human how to get $M^H$ closer to $M^R$ --What is the minimum number of changes needed in $M^H$ such that $\pi^R$ would be optimal plan.
Challenges in Human-Aware Planning

• Interpret what humans are doing based on incomplete human preference and domain models (Modeling)
  • Plan/goal/intent recognition

• Plan with incomplete domain models (Decision Making)
  • Robust planning/execution support with “lite” models
  • Proactive teaming support

• Explicable Behavior, Explanations/Excuses (Interaction/Communication)
  • How should the human and robot coordinate

• Understand effective interactions between humans and machines (Evaluation)
  • Human factor study
Overview of our work

How to learn and plan with incomplete domain models
Complete--Approximate--Shallow

How to plan to be useful to the human
Avoiding conflicts and offering serendipitous help

How to make planned behavior explicable or provide explanations to the human in the loop
Humans will parse the behavior in terms of their understanding of the Robot’s model

How to recognize and evaluate what are the desiderata for fluent teaming with humans

As the “paper clip” assistant shows, we AI’ers are not great at guessing what humans “like” 😐
Overview of our ongoing work

How to learn and plan with incomplete domain models

Complete--Approximate--Shallow

How to plan to be useful to the human
Avoiding conflicts and offering serendipitous help

How to make planned behavior explicable or provide explanations to the human in the loop
Humans will parse the behavior in terms of their understanding of the Robot’s model

How to recognize and evaluate what are the desiderata for fluent teaming with humans

As the “paper clip” assistant shows, we AI’ers are not great at guessing what humans “like”.
Spectrum of Domain Models

Increasing degree of incompleteness of planning models

No Model → Shallow Models → Partial Models → Approximate Models → Full Model

No plan → Plan critiquing or auto-completion → Planning Guidance → Robust plan generation and management → Traditional planning

Ease of learning/acquiring the models

Note the contrast to ML research where the progress is going from uninterpretable/non-causal models towards interpretable and causal models. So we might meet in the middle!
Action Vector Models

- View observed action sequences as “sentences” in a language whose “words” are the actions
- Apply skip-gram models to these sequences and embed the action “words” in a higher dimensional space
  - The proximity of the action words in that space is seen as their “affinity”
- Use the action affinities as a way to drive planning and plan recognition
Action Vector Models can be used to Recognize Plans

With the learnt vectors $w_j$, we can predict the target plan (as the most consistent with the affinities). We use an EM procedure to speedup the prediction.

$$\mathcal{F}(\hat{p}) = \sum_{k=1}^{M} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{k+j} | w_k)$$

- $M = |\text{the target plan}|$

The target plan to be recognized

Learning shallow models can avoid overfitting!!

Nominated for Best Student Paper Award at [AAMAS16]
Overview of our ongoing work

How to learn and plan with incomplete domain models

- Complete
- Approximate
- Shallow

How to plan to be useful to the human

- Avoiding conflicts and offering serendipitous help
- How to make planned behavior explicable or provide explanations to the human in the loop

Humans will parse the behavior in terms of their understanding of the Robot’s model

How to recognize and evaluate what are the desiderata for fluent teaming with humans

As the “paper clip” assistant that shows we Al’ers are not great at guessing what humans “like”...
Interaction Requires Modeling the Human

Explicability: Aim to get $\pi^R$ closer to $\pi^H$ (by getting $M^R$ closer to $M^H$)

Explanation:
Tell human how to get $M^H$ closer to $M^R$ --What is the minimum number of changes needed in $M^H$ such that $\pi^R$ would be optimal plan.
When is a plan “Explicable” to the human in the loop?

- The robot generates its plan of action using its model $M_R$
- The human “interprets” this plan in light of her understanding of the Robot’s model $M^*_R$
- $M_R$ and $M^*_R$ can be quite different.

Differences can be a result of:
- Different capabilities (e.g., possible actions)
- Different knowledge (e.g., level of modeling)
- Different interpretation of behaviors (e.g., plans) interacting with the world -- more than just trajectory planning!

$$\arg\min_{\pi_{M_R}} cost(\pi_{M_R}) + \alpha \cdot dist(\pi_{M_R}, \pi_{M^*_R})$$

But, alas, $M^*_R$ is not known!
Learning Human Expectation via Explicability Labeling

Analogy: Think of learning how to write address labels so the postal carrier can understand.

- Task labels (to associate with actions).
  - Collect
  - Store
  - Observe

More than one label is allowed for actions

\[
\arg\min_{\pi_{MR}} \text{cost}(\pi_{MR}) + \alpha \cdot \text{dist}(\pi_{MR}, \pi_{\hat{M}R})
\]

\[
\text{dist}(\pi_{MR}, \pi_{\hat{M}R}) = F \circ \mathcal{L}^*(\pi_{MR})
\]

\[
\arg\min_{\pi_{MR}} \text{cost}(\pi_{MR}) + \alpha \cdot F \circ \mathcal{L}_{\text{CRF}}^*(\pi_{MR} | \{S_i \mid S_i = \mathcal{L}^*(\pi_{MR})\})
\]
Web Interface to collect human feedback on robot task plans

- **Goal:** Create a web application that enables researchers to leverage crowd sourcing services (e.g., mechanical turker) to perform HRI studies in a simulated environment.

- **We are specifically interested in enabling users to annotate and/or modify robot task plans being presented to them.**

- **Related Projects:**
  - [http://jpdelacroix.com/simiam/](http://jpdelacroix.com/simiam/)
  - [http://robotwebtools.org/](http://robotwebtools.org/)
Bi-LSTM as Task Predictor for Plan Explicability

Motivation:
1. Consider future inputs.
2. Break Markov Property.

Feature:
- Action (0~N) + Executor (0-Human/1-Robot/2-Neither) + State (0010...)

Testing Accuracy: 90.76%

<table>
<thead>
<tr>
<th>noop</th>
<th>near-r1</th>
<th>at-b6-r1</th>
<th>at-b1-r1</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
Example 1 – Fetchworld

- Fetch needs to tuck its arms before moving

(move loc2 loc1)
(pickup b1 loc1)
(tuck)
(move loc1 loc2)
(putdown b1 loc2)

Explanation >> MOVE_LOC1_LOC2-has-precondition-HAND-TUCKED

(move loc2 loc1)
(pickup b1 loc1)
(move loc1 loc2)
(putdown b1 loc2)
Minimal Explanation (ME) vs Minimally Complete Explanation (MCE)

Robot Model

Human Model

(:init (block-at b1 loc1) (robot-at loc1) (hand-empty))

(:goal (and (block-at b1 loc2)))

(plan (pick-up b1) (tuck) (move loc1 loc2) (put-down b1))

"Beyond Explanations as Soliloquy"
IJCAI 2017
Trust in Autonomy

• One holy-grail in human aware AI systems is engendering trust in the humans
• The mechanisms of long term trust are complex
• However, ability of the agent to show *explicable behavior* and provide comprehensible *explanations* are clearly critical for engendering trust
• (Other factors: Assessment of self-competence and human competence)
Overview of our ongoing work

How to learn and plan with incomplete domain models
Complete—Approximate—Shallow

How to plan to be useful to the human
Avoiding conflicts and offering serendipitous help

How to make planned behavior explicable to the human in the loop

Humans will parse the behavior in terms of their understanding of the Robot’s model

How to recognize and evaluate what are the desiderata for fluent teaming with humans

As the “paper clip” assistant shows, we AI’ers are not great at guessing what humans “like” 😊
Do we really know what (sort of assistance) humans want?

Proactive Help Can be Disconcerting!
Human-human Teaming Analysis in Urban Search and Rescue

Simulated search task (Minecraft) with human playing role of USAR robot
- 20 internal/external dyads tested
- Conditions of autonomous/intelligent or remotely controlled robot
- Differences in SA, performance, and communications
Analysis of Proactive Support in Human-robot teaming
Simulated search task (Webots) with human remotely controlling a robot while collaborating with an intelligent robot ‘Mary’:

Findings
Robot with a proactive support capability (vs. without):
  - Higher dyad performance
  - Lower communication
  - Slightly (non-significant) increased mental workload

• Mary with a proactive support capability in our USAR task scenario is generally preferred

[IROS, 2015]
Summary

How to learn and plan with incomplete domain models

Complete—Approximate—Shallow

How to plan to be useful to the human

Avoiding conflicts and offering serendipitous help

How to make planned behavior explicable to the human in the loop

Humans will parse the behavior in terms of their understanding of the Robot’s model

How to recognize and evaluate what are the desiderata for fluent teaming with humans

As the “paper clip” assistant shows, we AI’ers are not great at guessing what humans “like”
Summary of the talk

• Part I: The Path to General AI goes through Human-Machine Collaboration
  • ...and it is a good thing!
    • Expands reach and scope of AI enterprise
    • Reduces some of the off-the-top worries about AI
    • Brings up novel research challenges

• Part II: Planning Challenges in Human-Machine Collaboration
  • Brief review of how the planning problem “expands” in the face of interaction/teaming with humans
  • Specific challenges and some ongoing work in my group
AI Challenges in Human–Robot Cognitive Teaming

Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, Yu Zhang

(Submitted on 15 Jul 2017)

Among the many anticipated roles for robots in future is that of being a human teammate. Aside from all the technological hurdles that have to be overcome on the hardware and control sides to make robots fit for work with humans, the added complication here is that humans have many conscious and subconscious expectations of their teammates -- indeed, teaming is mostly a cognitive rather than physical coordination activity. This focus on cognitive coordination, however, introduces new challenges for the robotics community that require fundamental changes to the traditional view of autonomous agents.

In this paper, we provide an analysis of the differences between traditional autonomous robots and robots that team with humans, identifying the necessary teaming capabilities that are largely missing from current robotic systems. We then focus on the important challenges that are unique and of particular importance to human–robot teaming, especially from the point of view of the deliberative process of the autonomous agent, and sketch potential ways to address them.

Subjects: Artificial Intelligence (cs.AI)

Cite as: arXiv:1707.04775 [cs.AI]
(or arXiv:1707.04775v1 [cs.AI] for this version)
Thank you