Trust, Expectation and Acceptance in Human-Automation Interaction

Liz Sonenberg

Department of Computing and Information Systems, University of Melbourne
l.sonenberg@unimelb.edu.au

EDTAS Presentation, July 2015

POSTERA CRESCAM LAUDE

THE UNIVERSITY OF
MELBOURNE

Trust – a multi-dimensional construct

"The human cognitive aspect of trust arises from our ability, based on various dimensions of commonality, to make reasonable inferences about the internal state of other actors (e.g., beliefs, dispositions, intentions) in order to predict future behaviour and judge the risk versus benefit of delegation.

... It is therefore crucial that agents not only correctly use the social interface, but also provide 'honest signals' about the agent's state, for a human partner to construct beliefs about the agent that accurately reflect its internal state." *

(Atkinson, 2012), *(Atkinson, 2013)
(Hancock et al, 2011), (Lee & See, 2004), (Wagner, 2013)
image: `https://www.youtube.com/watch?v=8osRaFTtgHo`

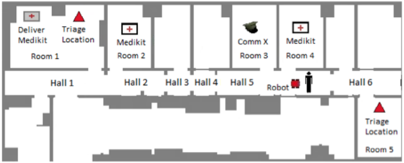
Foundations – Setting the Context

- Seeking to engender well-calibrated trustworthiness of automation by human participants in hybrid human-automation teams
- Scenarios involving goal-driven distributed interactions with multiple 'semi-autonomous' actors in complex dynamic settings
  - Varying degrees of interdependence and coupling of actions
  - Changing communication opportunities
  - Actors with diverse cognitive characteristics and varied awareness of others
- Need to design for ‘intrinsic and extrinsic cognitive capabilities’ (Lemaignan & Alami, 2014)
- For humans, trust guides reliance when complexity and unanticipated situations make a complete understanding of the automation impractical (Lee & See, 2004)

(Lee & See, 2004)
(Lemaignan & Alami, 2014)
(Zilberstein, 2015)

Foundations – Our Focus

- Deploy an agent's mental modelling of others' status and cognitive ability to influence interaction behaviour
- Exploit contemporary developments in automated planning



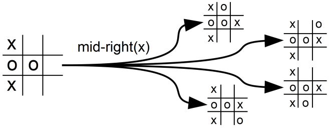
```
H: Comm. X is going to perform triage in room 5.
R: Okay.
H: I need you to bring a medical kit to room 1.
R: Okay.
```

```
H: I just put a new medical kit in room 4.
H: Comm. X is going to perform triage in room 5.
R: Okay.
H: I need you to bring a medical kit to room 1.
R: Okay.
```

(Felli et al, 2014), (Felli et al, 2015)
(Muise et al, 2014), (Muise et al, 2015)
image source: (Talamadupula, 2014)

Foundations – Key Observations

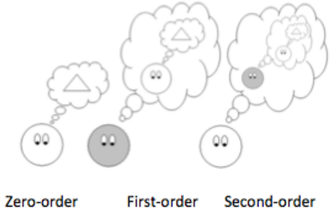
- Computational complexity considerations re brute-force approaches ...
- For a single agent operating in a multi-agent world, planning should not be omnipotent, but rather conditioned on what others may do
- Allow an agent not only to reason about others, but to assume their perspective
- Predict (and exploit) the effects that an agent's behaviour will have on the mental states of others
- Prune search via plausibility: if we know the goal of an agent, then we should only consider their plausible actions, given the context
- Multi-agent planning via translation to fully-observable non-deterministic single agent planning:



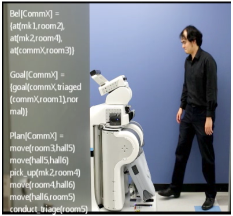
(Brafman & Domshlak, 2013)
(Muise et al, 2015)

Expectation – Setting the context

- A component of shared awareness is mutual predictability
  - failure in predictability results in expectation violation
  - constructing expectation violation is the core of surprise in magic tricks, but generally problematic in other circumstances
- In human-automation interaction, expectation breakdown occurs
  - via the psychological perspective (e.g. individual human approaches to interacting with physical robots) and
  - via the absence of requisite knowledge or beliefs
- Expectations play a role as meta-level computational constructs that impose reasoning obligations (*our focus)*
- Need to be able to manipulate beliefs about beliefs - i.e. *nested beliefs*
  - Do not assume all actors have the same 'cognitive' capability



Theory of Mind (Premack & Woodruff, 1978)

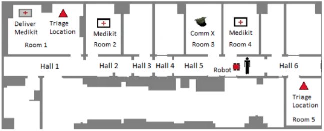


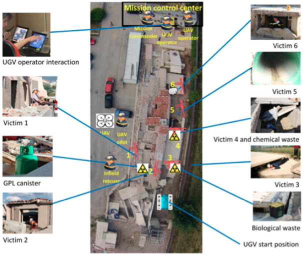
```
Bel[CommX] =
{at(mk1,room2),
at(mk2,room4),
at(commX,room3)}

Goal[CommX] =
{goal(commX,triaged
(commX,room1),nor
mal)}

Plan[CommX] =
move(room3,hall5)
move(hall5,hall6)
pick_up(mk2,room4)
move(room4,hall6)
move(hall6,room5)
conduct_triage(room5)
```

left image: tom.lisepijl.nl; right image: (Talamadupula, 2014), http://tinyurl.com/beliefs-anno

Expectation – Individual and Organisational

- *Interdependencies* between tasks, the task setting, and the actors influence the nature and timing of information to be shared
- The design of computational social structures (*teams*, *relationships*, and *organisations*) also shapes the timing and forms of delegation
- Our work (*so far only in simulation mode*) explores how awareness of those interdependencies can focus information exchange re confirmation or violation of expectations to achieve goals efficiently

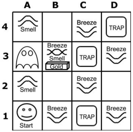




(Keogh et al, 2014), (Singh et al, 2014)
right image: (Kruijff, 2015; Figure 2)

Expectation – Computing with complex nested beliefs

- Our work supports both *stereotypical* and *empathetic* reasoning
    - empathetic – agent A simulates the reasoning that agent B would do from $B's$ perspective, with $B's$ belief base, and $B's$ cognitive apparatus as understood by A
- An agent can use one model for itself, and use different representations and reasoning mechanisms for others
- Notions of plausible and acceptable runs, based on the *social context* assumed by the agent
- Automated verification (*ATL model checking*) to compute strategies involving only acceptable runs
- Demonstrated in a *wumpus world* variant - a coordination game with added social features



	A	B	C	D
4	Smell		Breeze	TRAP
3		Breeze Smell Gold	TRAP	Breeze
2	Smell		Breeze	
1	Start	Breeze	TRAP	Breeze

(Felli et al, 2014), (Felli et al, 2015)
left image: `http://w3.sista.arizona.edu/classes/ista550/.../figs/wumpus-cave.jpeg`

Expectation – Computing with complex nested beliefs, ctd.

The lord of a castle is informed by a peasant that a wumpus is dwelling in a dungeon nearby. It is known that the wumpus can be killed by one hunter alone only if asleep; if awake, two hunters are required.

...The lord tasks the peasant to fetch the White Knight, his loyal champion, and hunt down the beast together. The White Knight is known for being innocent, trustworthy and brave; however, the peasant does not know any knight, and neither how they appear.

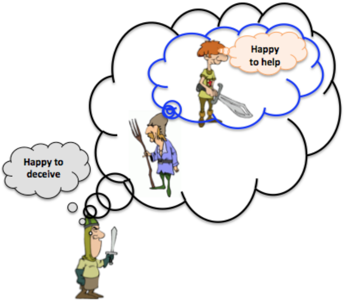
...While looking for the White Knight, the peasant runs into the Black Knight and, believing him to be the White Knight, tells him about the quest, which this knight accepts, willing to keep the gold. The Black Knight is aware of the misperception, but is happy to deceive the peasant.

Expectation – Computing with complex nested beliefs, ctd.

The lord of a castle is informed by a peasant that a wumpus is dwelling in a dungeon nearby. It is known that the wumpus can be killed by one hunter alone only if asleep; if awake, two hunters are required.

...The lord tasks the peasant to fetch the White Knight, his loyal champion, and hunt down the beast together. The White Knight is known for being innocent, trustworthy and brave; however, the peasant does not know any knight, and neither how they appear.

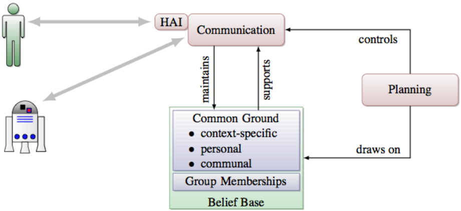
...While looking for the White Knight, the peasant runs into the Black Knight and, believing him to be the White Knight, tells him about the quest, which this knight accepts, willing to keep the gold. The Black Knight is aware of the misperception, but is happy to deceive the peasant.



(Felli et al, 2014), (Felli et al, 2015)
cartoons (CC-BY-SA) http://www.hasslefreeclipart.com/cart_fantasy/

Acceptance - establishing common ground

- *Common ground* is the information that participants in a joint activity share and assume to be shared. Establishing common ground is key to fluid interaction between actors in a joint activity.
  - Common ground can be defined via *belief* (Stalnaker, 2002), or by *acceptance* (Tuomela, 2003)
  - Social psychologists have identified distinctive components of common ground. We seek to represent and exploit them computationally, in conjunction with reasoning through *stereotypes*
- We have developed a modal logic characterisation of common ground as a further step towards precise analysis for mechanisms that use the concept

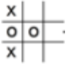


(Pfau et al, 2014), (Pfau et al, 2015)

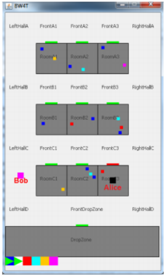
Acceptance – plausibility and heuristics

- We can demonstrate efficient agent decision making in a multi-agent context, exploiting modern planning techniques, in (gradually) increasingly complex settings

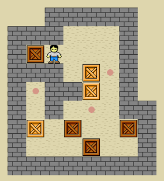
Tic-Tac-Toe



Blocks World for Teams



Sokoban



(Muise et al, 2014), (Muise et al, 2015)

Focused Challenges

So far we have:

- (Some) integration of planning with epistemic and doxastic concepts - ie taking account of other agents' knowledge and beliefs - and their beliefs of others' beliefs, etc
- (Some) incorporation of awareness of other agents' goals in an agent's planning and decision making

Followup work

- More powerful planning algorithms that are responsive to the needs of human users, and can accept human advice (c.f. *human aware task planning* (Lallement, 2014); *type II semi-autonomous systems* (Zilberstein, 2015))
- Interfaces that allow smooth transfer of control between automation and humans (supporting interleaved human-automation plan achievement), and that support inference at individual, team, and organisational levels
- Moving from generating shared awareness to repairing/restoring through social interaction - eg sharing information, adjusting control modes

General Challenges

- How can an automation component signal its intentions, status or 'personality' (e.g. risk profile)? including what to convey and how to achieve it through multi-modal interfaces?
- Are there sweet-spots in trading off design-time explicit knowledge with experience gathered and deployed at run-time?
- How much knowledge about human interpersonal trust is applicable to human interaction with automation?
- Murphy's 100:100 Challenge: can we achieve hundreds of remote knowledge workers independently directing and consuming information from hundreds of heterogeneous robots? (Murphy, 2011)
- Community-wide challenge problems for evaluation of distributed situation awareness as well as dynamic levels of autonomy, adapted to task, the situation, and participants' capabilities and needs. (International exemplars: ORCHID `www.orchid.ac.uk`, TRADR `www.tradr-project.eu`)

> *"The human cognitive aspect of trust arises from our ability, based on various dimensions of commonality, to make reasonable inferences about the internal state of other actors (e.g., beliefs, dispositions, intentions) in order to predict future behaviour and judge the risk versus benefit of delegation. … It is therefore crucial that agents not only correctly use the social interface, but also provide 'honest signals' about the agent's state, for a human partner to construct beliefs about the agent that accurately reflect its internal state."* (Atkinson, 2013)

Current collaborators and students

List of key references available at `http://tinyurl.com/EDTAS-LizS`