ihmc

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Shared Awareness, Autonomy and Trust in Human-Robot Teamwork

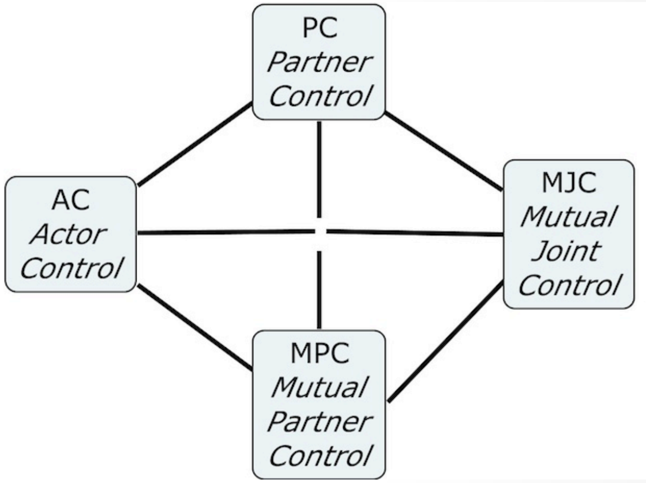
David J. Atkinson, William J. Clancey, and Micah H. Clark
Institute for Human and Machine Cognition

Presented at AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction
Arlington, VA 13-15 November 2014

The Theory

- **Effective teamwork**
  - ➔ **mutual trust**
    - ➔ **shared awareness**
      - ➔ **aligned mental models**
        - ➔ **expectations**
  - Actors, activities, situations
  - What has happened in the past and why; what is happening now

- **Failed expectations**
  - ➔ **loss of trust**
    - ➔ **explanations** + ***remedies***
      - ➔ **repaired trust**



PC
Partner Control

AC
Actor Control

MJC
Mutual Joint Control

MPC
Mutual Partner Control

Control Authority
& Interdependency

A key remedy to repair trust is **adaptive autonomy**

The Theory in One Slide

- Effective teamwork requires **mutual trust**
- Establishment and maintenance of mutual trust requires **shared awareness**
- Shared awareness requires continual alignment of **mental models**
  - Actors, activities, situations
  - What has happened in the past and why; what is happening now
- Mental models serve as a source of **expectations**
- When expectations fail, **mutual trust may fail**
- Trust is maintained when failed expectations are **explained**, and **remedies are applied**
- A key remedy is **adaptive autonomy**

Expectation Violations

- A failure of **predictability:** an inconsistency between the *expected* and *actual* state of the world as perceived by human and/or robot
  - **Unilateral** (one actor) or **Bilateral** (both actors)
- **Explanations**: identification of the source of divergence in shared awareness (mental models)
  - Attribution to belief(s) about the other team member, about other agents, exogenous conditions, the task at hand …
- **Choice of method for restoring** shared awareness
  - **Explanations**, relative justification of **beliefs**, symmetry of **information**, assessment of **potential outcomes**
- Effective repair requires **social interaction** between robot and human to adjust beliefs, task, methods

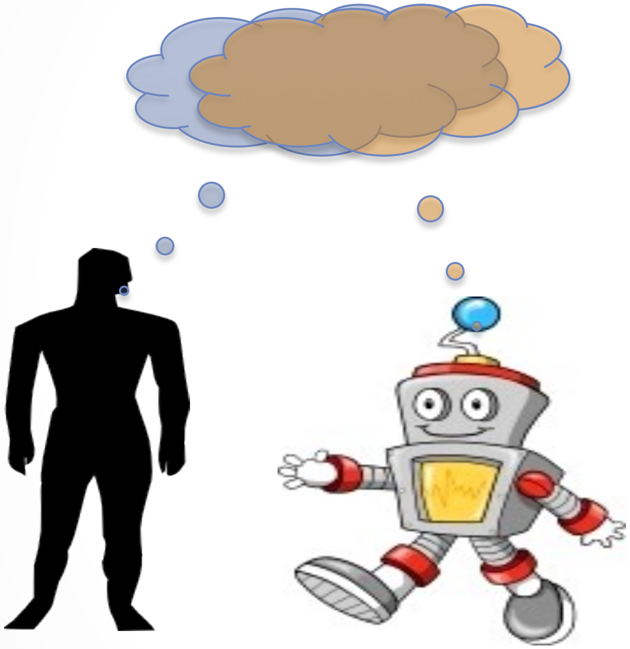
Adaptive Autonomy

- Refers to (unilateral) action by a robot to achieve team goals with fluid changes in interdependency
  - Dynamic change in control modes *at multiple levels of abstraction* and *instantiation* within a system

- Change/adaptation occurs along three dimensions
  - **Commitment:** Range of implicit to explicit delegation/acceptance of task
  - **Specification:** Range of task description from abstract to concrete
  - **Control Authority:** interdependency states and transitions defined by relative mutual or joint control of outcomes, scope of independent action, degree of symmetry in access to important information

- A robot adjusts autonomy by invoking actions that lead to control model state transitions
  - Restoration of **shared awareness** and predictability

Thank You



datkinson@ihmc.us