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 (IHMC)
15 SE Osceola Ave., Ocala, FL 34471
{datkinson, wclancey, mclark} @ihmc.us

Abstract

Teamwork requires mutual trust among team members. Establishing and maintaining trust depends upon alignment of mental models, an aspect of shared awareness. We present a theory of how maintenance of model alignment is integral to fluid changes in relative control authority (i.e., adaptive autonomy) in human-robot teamwork.

Extended Abstract

The foundation of teamwork is well-calibrated mutual trust among team members. The goal of our research is to enable trust for appropriate reliance and interdependency in teams composed of humans and robots: such teams may be found in any application domain that requires coordinated joint activity by humans and intelligent agents, whether those agents are embedded in cyber-physical systems (e.g., air traffic control; dock yard logistics) or embodied in robots (e.g., robots for assisted living; a surgical assistant). We hypothesize that establishing and maintaining trust depends upon alignment of mental models, which is at the core of team member shared awareness. Secondly, maintaining model alignment is integral to fluid changes in relative control authority (i.e., autonomy) as joint activity unfolds.

Team members are engaged in parallel, distributed actions whose interactions may be synchronous or asynchronous, with various degrees of interdependence and information exchange, and actions may only be loosely coupled. A dynamic and uncertain environment compounded with the complexities of coordinated teamwork may lead to unexpected effects for each team member, including loss in shared awareness. Accomplishing tasks will involve resolution of conflicts among numerous interacting factors, and this may require a dynamic response by the team. It is in this environment we find the greatest challenges to maintaining mutual trust among human team members. Responding to perturbations that endanger trust is crucial for optimal human teamwork; we believe that similar challenges are present for human-robot teams.

Unlike traditional automation, robotic autonomous agents may resemble human teammates: they may have discretion in what they do, and their need for supervision may vary. Like humans, they may differ in competence, adapting to the unknown, and self-knowledge. Autonomous agents are in fact *actors*. Autonomy is not only the ability to independently perform actions, but to choose what goals to pursue and in what manner; to volunteer; and to take or concede the initiative when needed. Teamwork between person and agent requires interdependence, coordination, and cooperation, implying well-structured interactions to establish these states and fluid changes in control authority.

We assert that successful team interaction and changes in control require shared understanding, e.g., of actors, activities, and situations. All are components of *shared awareness*, shown previously to strongly affect trust among human teammates (Muir 1994). "Common ground" also reduces the communication required to coordinate action (Kiesler 2005).

Shared awareness, a product of what has happened in the past and what is happening now, is a dynamic, continually refreshed and resynchronized source of mutual team member *expectations*, including evolution of team member interdependencies, individual behavior, task activities, and situational factors. For example consider a carpenter's expectation that his workmate will hold a board firmly while he nails it in place. Explicit model-based expectations, when based on context-sensitive projection of plans, have proven in non-teamwork applications to be a powerful tool for focusing attention, verifying, monitoring, and controlling complex systems (Atkinson and James 1990). Our research seeks to extend expectation-based monitoring and control to coordinated human-robot teamwork. We also

build upon studies of teamwork that link successful coordination to expectations of each partner's actions (Knoblich and Jordan 2003) and show that anticipatory robot actions based upon expectations about human collaborators give rise to a perception of "fluency" of robot action and *predictability* (Hoffman and Breazeal 2007).

Predictability is at the core of belief that a desired outcome, to be brought about by a trusted agent, will occur (Golembiewski and McConkie 1975). The attribution of predictability has been shown to be especially important for trust in automation (Atkinson and Clark 2014), (Marble et. al. 2004), (Muir 1994). To achieve predictability, a robot requires a rich representational system to support theories of mind and an ability to project these models into the future. Acting on these projections builds predictability, shaping the person's model of the agent. Our approach uses the Brahms multi-agent simulation framework (Clancey et. al. 1998), (Clancey 2002), and the ViewGen system (Ballim and Wilks 1990).

A failure of predictability results in an *expectation violation*: an inconsistency between the expected and actual state of the world as perceived by human or robot. Such violations are a cause of breakdowns in teamwork. *Bilateral expectation violations* occur when the expectations of both actors fail. This type of violation can often be resolved via information gathering; the cause is likely external to the team, e.g., an un-modeled change in the environment.

A *unilateral expectation violation* occurs when the expectation of only one of the actors fails. This may be due to unexpected omission/commission of control actions by a teammate (e.g., the carpenter's helper releases the board before the final nail is in place, causing it to be mispositioned). This is of greater concern because it reflects a *divergence in shared awareness*. If left uncorrected, such a violation threatens predictability and therefore mutual trust.

To recover predictability, an *explanation* of an expectation violation is required. When a team member's competence is uncertain, the reliability of their ability to contribute to shared goals becomes compromised. Failure by a robotic agent to notice such an attribution by a human teammate, or to respond appropriately, may lead to catastrophic loss of trust in the robot.

Restoring shared awareness through social interaction (Atkinson and Clark 2013) is crucial in resolving an expectation violation. Remedies may include modifying shared beliefs, realigning models or changing control authority or tasks. The choice of repair method depends upon the violation's source attribution (one or both actors, or the situation), the justification of beliefs at the basis of the expectation, and symmetry of information access by team members. For example, the carpenter's robot assistant might explain that it thought two nails would be sufficient and didn't expect the board to drop. Rapid explanation and acceptance of responsibility (if indicated) helps restore trust (Lewicki and Wiethoff 2000). Another remedy is modifying relative control authority (aka adaptive autonomy). Changing control authority may tradeoff task optimality for increased trust (e.g., requesting step-by-step guidance).

We view robot autonomy as a *multi-dimensional characteristic of control modes* for carrying out a particular activity *within* the context of other activities and external situation. Adaptive autonomy is highly dynamic; even in the normal course of task achievement joint activities may have different control modes at different levels of abstraction and instantiation. Control modes reflect the complexity of interdependency between human and robot teammates.

Our theory defines control modes and provides for adaptation along three principal dimensions of autonomy: *Commitment*, *Specification*, and *Control.* A change along the *Commitment* dimension affects shared awareness by increasingly explicit task delegation or acceptance where dependency may have heretofore been implied. Intervention along the *Specification* dimension may represent a change in the degree of "help" provided. Specification changes may entail a corresponding change in the *Control* dimension, which adjusts interdependency by transitioning among situational states that define relative joint control of outcomes, independence of control actions, etc.

In our approach, the robot agent adjusts autonomy by invoking actions that lead to a target state transition, where the target transition is a function of (1) the explanation of the expectation violation; (2) justified differences in shared awareness, (3) degree of symmetry in access to task-control information, and (4) impact on trust or achievement of desirable outcomes. Actions adjusting autonomy ought to include social interaction to communicate the rationale. Transitions to high robot autonomy are not likely to be abrupt except in cases of *bona fide* emergencies. Crucially, a robot requires a degree of self-knowledge to take initiative in changing control authority, and this bar is highest when it is towards a state of greater autonomy.

We suggest that the greater the extent of shared awareness among human and robot team members, the greater mutual trust and the likelihood that structured social interactions will fluently achieve successful transitions in control authority—the essence of well-coordinated teamwork.

References

Atkinson, D.J. and Clark, M.H. 2013. Autonomous Agents and Human Interpersonal Trust: Can We Engineer a Human- Machine Social Interface for Trust? *In Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium. Technical Report No. SS-13-07.* Menlo Park, CA: AAAI Press.

Atkinson, D.J. and Clark, M.H. 2014. Attitudes and Personality in Trust of Intelligent, Autonomous Agents. Submitted manuscript.

Atkinson, D. and James, M. 1990. Applications of AI for automated monitoring: The SHARP system. In Proceedings of the AIAA Second International Symposium on Space Information Systems. Pasadena, CA: American Institute of Aeronautics and Astronautics.

Ballim A. and Wilks, Y. 1991. *Artificial Believers: The Ascription of Belief.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Clancey, W.J., Sachs, P., Sierhuis, M., van Hoof, R. 1998. Brahms: Simulating Practice for Work Systems Design. *International Journal on Human-Computer Studies* 49:831–865.

Clancey, W.J. 2002. Simulating Activities: Relating Motives, Deliberation, and Attentive Coordination. *Cognitive Systems Research* 3(3):471–499.

Golembiewski, R.T. and McConkie, M. 1975. The Centrality of Interpersonal Trust. Cooper, C.L ed. *Theories of Group Processes.* Australia: John Wiley & Sons.

Hoffman, G. and Breazeal, C. 2007. Effects of Anticipatory Action on Human-Robot Teamwork. In *Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction* 1-8. New York, NY: ACM Press.

Kiesler, S. 2005. Fostering common ground in human-robot interaction. In *Proceedings of IEEE International Workshop on Robot and Human Interactive Communication.* 729-734. Pittsburgh, PA: IEEE Press.

Knoblich, G. and Jordan, J.S. 2003. Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(5):1006–1016.

Lewicki, R.J. and Wiethoff, C. 2000. Trust, Trust Development, and Trust Repair. M. Deutsch & P.T. Coleman. eds. *The handbook of conflict resolution: Theory and practice.* 86-107. San Francisco, CA: Jossey-Bassa.

Marble J, Bruemmer D, Few D, and Dudenhoeffer D. 2004. Evaluation of Supervisory vs. Peer- Peer Interaction with Human-Robot Teams. In *Proceedings of the 37th Hawaii International Conference on System Sciences.* Hawaii: IEEE Press.

Muir B. M. 1994. Trust in automation: Part I. Theoreticalissues in the study of trust and human intervention in automated systems. *Ergonomics* 37(11):1905–1922.