Curing Robot Autism: A Challenge

Gal A. Kaminka
The MAVERICK Group
Computer Science Dept. and Gonda Brain Research Center
Bar Ilan University, Israel
galk@cs.biu.ac.il

ABSTRACT

Almost all robots are autistic; very few humans are. Out of the box, robots generally do not behave correctly in social settings (involving humans, or other agents). Most researchers treat this challenge *behaviorally*, by superficially tacking task- and domain- specific social behavior onto functioning individual robots. These rules are built once, and applied once. In contrast, I posit that we can build better socially-capable robots by relying on general social intelligence building blocks, built into the brains of robots, rather than grafted on per mission: *built once, applied everywhere*. I challenge the autonomous agents community to synthesize the computational building blocks underlying social intelligence, and to apply them in concrete robot and agent systems. I argue that our field is in a unique position to do this, in that our community intersects with computer science, behavioral and social sciences, robotics, and neuro-science. Thus we can bring to bear a breadth of knowledge and understanding which cannot be matched in other related fields. To lend credibility for our ability to carry out this challenge, I will demonstrate that we have carried out similar tasks in the past (though at a smaller scale). I conclude with a sample of some open questions for research, raised by this challenge.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]; I.2.9 [**Robotics**]

Keywords

Challenge; Robot Autism; Social Intelligence; Multi-agent systems; Multi-robot systems

1. ROBOT AUTISM AND ITS IMPACT

Robots and other synthetic agents (e.g., virtual humans) are generally autistic. Out of the box, before a significant programming effort is spent, they are unable to behave correctly in social settings, involving other agents, robots, or humans. And yet in a sense, making agents or robots behave correctly towards others (whether synthetic or human) is at the core of multi-agent and multi-robot systems research.

Most robotics researchers treat this challenge *behaviorally*, by superficially tacking task-specific social behavior onto functioning individual robots. The behavioral approach focuses on adding ad-hoc communication and coordination rules that will cover social behavior for the task at hand. Essentially, this entails narrowing the scope of the interactions, and spending very significant and expensive programming efforts, until the robots do *well enough*, in terms of their social behavior, at the task at hand. The analogy is to treatment of high-functioning autism by behavioral therapy, whereby the patient (in our case, the robot) is given rules to follow blindly to improve social functioning, without real internalization of their rationale. Examples abound in multi-robot systems. Roboticists typically address canonical tasks—one at a time—and general techniques emerge of studying those. But this methodology encourages this behavioral approach, where the interactions between robots are highly optimized for each specific task. The social behavior rules thus designed are built per-task, and are applicable only to the task.

In contrast, I posit that we can build socially-capable robots by developing general individual social intelligence building blocks, to be built into the brains of robots, rather than grafted on per mission: *built once, applied everywhere*. These building blocks, properly wired, would then cause the robots to behave socially, regardless of the task they are carrying out. The analogy to is to normal human minds: If we build our robots such that they posses even a subset of adult human social mechanisms (e.g., theory of mind, communications, collaboration, etc.), we would have robots that correctly function socially across a wide variety of tasks.

The advantages to such an approach would be significant, both to the science as well as to the engineering of social autonomous agents (and in particular robots). An understanding of hypothesized general social intelligence building blocks, would directly impact the directions the autonomous agents and multi-agent systems (AAMAS) community takes (focusing on general mechanisms and their parameterization, rather than a fragmented set of specialized problems and solutions). General social intelligence, essentially out-of-the-box, would very significantly reduce the prohibitive programming efforts currently associated with the design and deployment of commercial multi-robot systems. And it even carries the potential for impacting other fields, such as anthropology, cognitive science, and social (neuro-) science, on computational grounds.

I therefore challenge the autonomous agents and multi-agent systems community to synthesize the computational building blocks underlying social intelligence, to explore theories of social intelligence, and to validate them empirically in concrete robot and agent systems. I argue that our field is in a unique position to do this, in that our community intersects with computer science, behavioral and social sciences, robotics, and neuro-science. Thus we can bring to bear a breadth of knowledge and understanding which cannot be matched in other related fields.

To lend credibility for our ability to carry out this challenge, and to the benefits of succeeding at it, I will demonstrate that we have carried out at least one similar task in the past (though at a smaller

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.*

scale). In particular, I will briefly discuss previous AAMAS work in teamwork, which bears many similarities to the challenge at hand. Our success in investigating teamwork should inspire us to take on the larger-scale challenge presented here.

To start us off, I present in this paper an abstract hypothesis that general social intelligence can be modeled from four interconnected component processes: (1) recognizing other agents; (2) understanding their internal state; (3) measuring differences and similarities to the other agents; and (4) acting on the differences and similarities found. I conclude with a short sample list of open questions raised by this hypothesis.

2. GENERAL MECHANISMS FOR SOCIAL INTELLIGENCE

Researchers in artificial intelligence, cognitive science, autonomous agents, and multi-agent systems have been investigating components of social cognition for many years. A very partial list includes: recognizing intentionality in action (e.g., [21, 26, 4]), imitation [38, 3], theory of mind [39, 15, 35, 16, 37, 2], intent or plan recognition (e.g., [33, 7, 44, 30, 17, 1]), planning and execution in teams (e.g., [19, 27, 28, 9, 18, 42, 46, 40, 32]), observation-based coordination (e.g., [11, 25]), reasoning about conventions, commitments, norms, institutions, and organizations (e.g., [27, 24, 22, 23]), and many more.

However, with the exception of early theoretical work [6], we have for the most part avoided integrating these components to create a comprehensive picture of social intelligence. We (very) partially understand some of the components; we do not at all have a good picture of the system. This is the crux of the challenge.

To motivate researchers to address this challenge, I make the following hypothesis, in hopes that many researchers will rush to prove it insufficiently accurate. I hypothesize that *social intelligence is composed of the following component processes, that interact to generate social behavior*:

Recognizing Other Agents. One process identifies other agents in the environment (recognizing their existence as agents, rather than inanimate objects). Some cognitive scientists (e.g., [16, 21, 26, 2]) argue that humans employ perception inference rules to determine whether something is an agent (e.g., agents have self-propulsion and demonstrate goal-directedness; gaze also seems to be an indicator). Computational models can be built to allow a robot to distinguish agents from inanimate objects (this, as it turns out, is not only beneficial from an evolutionary perspective, but also computationally eases many robotic tasks, such as localizing).

Fundamental Mechanisms for Understanding Others. Once a robot recognizes that another agent is present, it can begin to predict its actions and intents (from an evolutionary perspective, a prudent step). There are two fundamental ways to gather information about the other: from communications, and by inference from observations. Most literature in our field focuses on the use of communications. This requires the other robot to cooperate in exchanging messages (or at least, in sending them).

The other method relies on observations. In AI, we typically refer to this as *plan recognition*. Almost all works in this area utilize a plan library to carry out the recognition, where the plan library is either learned, or manually constructed by the developer. But in 1995, agents researchers Tambe and Rosenbloom [44] have presented a method for agents in a simulated world to understand the observed actions of others—by explaining the observations in terms of the observer's own action repertoire. So instead of a specially-constructed recognition library, the agent would use its own planning knowledge to explain observed actions. They argued that from a computational standpoint, it was a cheap way for an agent to understand what others were doing.

Independently—and later—mirror neurons were found in primate (later, human) brains [41] that did exactly this: The neurons would fire when the primate was taking a specific action, or when it saw someone else take the same action; in other words, the neurons signified understanding others' actions using the observer's own knowledge of executing the actions.

I join others (e.g., [15]) in hypothesizing that this is a fundamental strategy to understanding others, i.e., to the capacity for theory-of-mind in agents. Moreover, I argue that this process occurs at various levels (simple actions to plans to goals and intentions). The robot recognizes the actions and intents of others in its own terms. It can thus represent others using the same data structures as it uses to represent its own executing control processes. Just as it knows its own actions, it knows those of the others; just as it knows its owns goals, it can infer the goals of others.

Comparing Myself to Others. A significant benefit of having a robot use the same knowledge base to describe its own actions, plans, and goals, as it does those of others, is that this facilitates comparison. By a process of continuous comparison of its own control structures and state to those of its peers, it can identify diverging decisions and status. This can indicate failures, if we have expectations that the robots have identical actions (or plans, or goals), as in teamwork [43, 31], or it can be used as a source of information on what actions (or plans, or goals) the observing robot should adapt (e.g., in imitation [14, 38, 3], or crowds [13]). The process is reminiscent of Social Comparison Theory [12], a social psychology theory.

Note that I mean comparison in the broad sense that includes measurement of generalized distance between states, not just equality testing. I also note that comparison does not imply any cooperation from the other side, nor does it exclude it. For example, in predatory settings, the predator might compare its own position and heading to those of its prey, without the prey's knowledge. Or in other words, the predator may take the distance between itself and the prey into account in deciding on its own next action.

Acting on the Generalized Distance to Others. Finally, the robot can act on the results of the comparison, the distance measurement. Given a complex state, some attributes may be similar (close), others may be different (far). Either can trigger actions.

Some can be in form of avoidance and negation, e.g., in avoiding collisions, or making a decision to take on a task that has not yet been allocated to a team-member (i.e., which involves *avoiding* to take on a task that has been taken by another). Others can be in the form of actions that minimize the differences between the robot and the object of comparison, as in mimicry or imitation [36], or in contagion processes [14, 45]. For instance, in teamwork literature [9, 19, 28, 43] agents take proactive steps to maintain agreement on joint goals and plans for execution by the team.

I stress that the process of acting on the differences and similarities may take place at different levels, from mimicry of actions, through imitation and internalization of plans (learning from demonstration [3]), to adopting the goals of another robot, potentially planning and executing a novel plan for this goal.

3. TEAMWORK: EVIDENCE THAT THIS CHALLENGE IS FEASIBLE

The challenge I am posing is a significant one. I therefore want to provide evidence that it can be tackled successfully by the AAMAS community. To do this, I will refer to a similar line of research,

though on smaller scale, that was successfully addressed by the AAMAS community in years past.

The argument I laid for why we should determine general social intelligence mechanism borrows from a similar argument made in the mid-1980's by AI and AAMAS researchers arguing for a general theory of teamwork, and by later arguments, on practical grounds, for the use of general teamwork software architectures to support construction of robust multi-agent systems. I briefly trace these investigations below.

In the mid-1980's, AI and AAMAS researchers inspired by studies of human teamwork (see [5] for a summary) have argued that teamwork should be treated as a domain of study, and that it can be formally described. Perhaps the first researchers to do this were Gross and Sidner[19], who proposed a formal method called SharedPlans for agents to reason about teams, and about their teammates (the SharedPlans model was later extended and improved in Grosz and Kraus [18]). Cohen and Levesque proposed the Joint Intentions Framework [9], an alternative model for proscribing teammate behavior. The Australian AI institute conducted additional influential theoretical investigations at the time [34].

This theoretical work was picked up and inspired researchers who have been becoming increasingly frustrated with the brittleness of distributed multi-agent systems, for industrial tasks [27, 28] and virtual training environments [43]. Essentially independently from each other, both Jennings and Tambe made arguments strikingly similar to the arguments I made above: That it is impractical, if not impossible, to continuously patch distributed multi-agent systems with more and more task-specific coordination rules. That instead, general teamwork mechanisms can and should be built for automating the collaborative interactions between the agents, in whatever domain of application the agents are used. Jennings proposed the GRATE* system to do this, extending the Joint Intentions Frameworks to cover commitments to joint plans, in addition to joint goals. Tambe proposed STEAM (a Shell for Teamwork), which went further in that it also possessed mechanisms for hierarchical teams and hierarchical tasks, decision-theoretic protocols for exchanging messages, and more. STEAM was inspired by both Joint Intentions as well as SharedPlans.

These initial investigations as to the practical benefits of using general teamwork mechanisms led to a series of improvements and applications, resulting in further practical impact (e.g., [40, 47, 10, 42, 29, 20, 32]), and theoretical extensions (some of which led to today's distributed POMDP investigations). Today, there exist even commercial software products that are built around the idea of automated general teamwork [8]. Of course, research in teamwork is continuing. And yet it is clear already that the idea of general teamwork proved very effective, both for AAMAS science as well as for its applications.

I emphasize that this brief review of the literature is meant only to argue for the possibility of success, and for the significance of impact, of the endeavor I am challenging the community to take.

4. LET'S GO!

I outlined an abstract hypothesis as to the key components of general social intelligence. But to make the challenge concrete, it would make sense to propose a specific set of milestones whose achievement would signify progress; or lacking specific milestones, at least a set of open problems. Also, just as teamwork research progressed by an interplay between theoretical progress and its application to concrete applications, so would research into general social intelligence mechanisms flourish in the presence of both concrete theories and concrete applications.

Two immediate—yet broad—directions for research are raised by the hypothesis above. One takes a *component view*, seeking to fill-in the details of each component above with hard-core computer science; to provide computational models and algorithms. The other takes a *system view*, seeking to argue for—or against—these components, and for—or against—specific ways in which they interact. The two directions are tied together, and impact each other.

There are of course many open questions, whose answers would shed light on the hypothesis above, both at the system view and component view. I cannot hope to list them all, yet I can point out a few candidates for immediate research:

- Can we show that good teamwork, whether described in logic [18] or in algorithms [43, 42, 32], is a special case of the abstract system described above, i.e., four components (specific instances) and specific interactions between them?
- Can we model and simulate additional types of group activities, known in social sciences, such as crowds [13], or treatment groups [20]?
- What is the role of learning, and how it is to be integrated in each component, and at the system level? Learning permeates social reasoning in humans, in many different ways; are there any specific forms that are better or worse for social reasoning in robots?
- Some specific instances (such as for imitation) look almost like planning: We are given an initial state (our own), and a goal state (that of the other agent), and the task is to determine actions to take us from the initial state to the goal state. Is this a special case of AI planning, applied to a domain that concerns social state? What are the operators, in this case?

Acknowledgments. Supported in part by ISF Grant #1511/12. I wish to thank the anonymous reviewers for their insightful comments. I enthusiastically thank Harvard University's Radcliffe Institute for Advanced Study for providing me with the space and time to crystallize the challenges outlined, and to begin pondering their resolution. I had received extremely useful comments and important critique from Barbara Grosz, Stephen Mann, John Plotz, Benny Shilo, Laurel Bossen, Ralph Hanna, Susanne Freidberg, Ray Jayawardhana, and Diane McWhorter. I thank them all, wholeheartedly. Lastly, I am indebted, beyond description, to Michael Fischer, Susann Wilkinson, and K. Ushi: Couldn't have written this without you.

5. REFERENCES

[1] D. Avrahami-Zilberbrand and G. A. Kaminka. Incorporating observer biases in keyhole plan recognition (efficiently!). In *AAAI-07*, pages 944–949, 2007.

[2] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1995.

[3] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In B. Siciliano and O. Khatib, editors, *Springer Handbook of Robotics*, pages 1371–1394. Springer, 2008.

[4] E. Bonchek-Dokow, G. A. Kaminka, and C. Domshlak. Distinguishing between intentional and unintentional sequences of actions. In *International Conference on Cognitive Modeling (ICCM-09)*, 2009.

[5] J. J. Burns, E. Salas, and J. A. Cannon-Bowers. Team training, mental models, and the team model trainer. In *Advancements in Integrated Delivery Technologies*, Denver, CO, 1993.

[6] K. Carley and A. Newell. On the nature of the social agent. *Journal of Mathematical Sociology*, 1994.

[7] E. Charniak and R. P. Goldman. A Bayesian model of plan recognition. *AIJ*, 64(1):53–79, 1993.

[8] CogniTeam, Ltd. CogniTAO (Think As One). Think As One.

[9] P. R. Cohen and H. J. Levesque. Teamwork. *Nous*, 35, 1991.

[10] T. D.Vu, J. Go, G. A. Kaminka, M. M. Veloso, and B. Browning. MONAD: A flexible architecture for multi-agent control. In *AAMAS-03*, 2003.

[11] M. Fenster, S. Kraus, and J. S. Rosenschein. Coordination without communication: Experimental validation of focal point techniques. In *ICMAS-95*, pages 102–108, California, USA, 1995.

[12] L. Festinger. A theory of social comparison processes. *Human Relations*, 7:117–140, 1954.

[13] N. Fridman and G. A. Kaminka. Modeling pedestrian crowd behavior based on a cognitive model of social comparison theory. *Computational and Mathematical Organizational Theory*, 16(4):348–372, 2010. Special issue on Social Simulation from the Perspective of Artificial Intelligence.

[14] N. Fridman and G. A. Kaminka. Towards a computational model of social comparison: Some implications for the cognitive architecture. *Cognitive Systems Research*, 12(2):186–197, 2011.

[15] V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences*, 2:493–501, 1998.

[16] G. Gergely and G. Csibra. Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Science*, 7(7), 2003.

[17] R. P. Goldman, C. W. Geib, and C. A. Miller. A new model of plan recognition. In *UAI-1999*, Stockholm, Sweden, 1999.

[18] B. J. Grosz and S. Kraus. Collaborative plans for complex group actions. *AIJ*, 86:269–358, 1996.

[19] B. J. Grosz and C. L. Sidner. Plans for discourse. In P. R. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 417–445. MIT Press, Cambridge, MA, 1990.

[20] M. Hadad, G. Armon-Kest, G. A. Kaminka, and S. Kraus. Supporting collaborative activity. In *AAAI-05*, 2005.

[21] K. Harui, N. Oka, and Y. Yamada. Distinguishing intentional actions from accidental actions. In *Proceedings of the 4th IEEE international conference on development and learning*, 2005.

[22] H. Hexmoor, S. G. Venkata, and D. Hayes. Modeling social norms in multiagent systems. *Journal of Experimental and Theoretical Artificial Intelligence*, 18(1):49–71, 2006.

[23] C. D. Hollander and A. S. Wu. The current state of normative agent-based systems. *Journal of Artificial Societies and Social Simulation*, 14(2), 2011.

[24] B. Horling and V. Lesser. A Survey of Multi-Agent Organizational Paradigms. *The Knowledge Engineering Review*, 19(4):281–316, 2005.

[25] M. J. Huber and E. H. Durfee. Deciding when to commit to action during observation-based coordination. In *ICMAS-95*, pages 163–170, 1995.

[26] S. Itakura, H. Ishida, T. Kanda, Y. Shimada, H. Ishiguro, and K. Lee. How to build an intentional android: infants' imitation of a robot's goal-directed actions. *Infancy*, 13(5), September 2008.

[27] N. R. Jennings. Commitments and conventions: the foundations of coordination in multi-agent systems. *Knowledge Engineering Review*, 8(3):223–250, 1993.

[28] N. R. Jennings. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *AIJ*, 75(2):195–240, 1995.

[29] G. A. Kaminka and I. Frenkel. Flexible teamwork in behavior-based robots. In *AAAI-05*, 2005.

[30] G. A. Kaminka, D. V. Pynadath, and M. Tambe. Monitoring teams by overhearing: A multi-agent plan recognition approach. *Journal of Artificial Intelligence Research*, 17:83–135, 2002.

[31] G. A. Kaminka and M. Tambe. Robust multi-agent teams via socially-attentive monitoring. *JAIR*, 12:105–147, 2000.

[32] G. A. Kaminka, A. Yakir, D. Erusalimchik, and N. Cohen-Nov. Towards collaborative task and team maintenance. In *AAMAS-07*, 2007.

[33] H. A. Kautz and J. F. Allen. Generalized plan recognition. In *AAAI-86*, pages 32–37. AAAI press, 1986.

[34] D. Kinny, M. Ljungberg, A. Rao, E. Sonenberg, G. Tidhar, and E. Werner. Planned team activity. In C. Castelfranchi and E. Werner, editors, *Artificial Social Systems, Lecture notes in AI 830*, pages 227–256. Springer Verlag, New York, 1992.

[35] A. N. Meltzoff. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5), 1995.

[36] A. N. Meltzoff. Imitation as a mechanism of social cognition: origins of empathy, theory of mind, and representation of action. In U. Goswami, editor, *Blackwellï¿½s Handbook of Childhood Cognitive Development*. Blackwell, 2002.

[37] A. N. Meltzoff. The "like-me" framework for recognizing and becoming an intentional agent. *Acta psychologica*, 2007.

[38] C. L. Nehaniv and K. Dautenhahn, editors. *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge University Press, 2007.

[39] D. G. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1, 1978.

[40] D. V. Pynadath and M. Tambe. Automated teamwork among heterogeneous software agents and humans. *JAAMAS*, 7:71–100, 2003.

[41] G. Rizzolatti, L. Fogassi, and V. Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature reviews in neuroscience*, 2001.

[42] P. Scerri, L. Johnson, D. Pynadath, P. Rosenbloom, M. Si, N. Schurr, and M. Tambe. A prototype infrastructure for distributed robot-agent-person teams. In *AAMAS-03*, 2003.

[43] M. Tambe. Towards flexible teamwork. *JAIR*, 7:83–124, 1997.

[44] M. Tambe and P. S. Rosenbloom. RESC: An approach to agent tracking in a real-time, dynamic environment. In *IJCAI-95*, 1995.

[45] J. Tsai, E. Bowring, S. Marsella, and M. Tambe. Empirical evaluation of computational emotional contagion models. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, 2011.

[46] D. E. Wilkins, T. Lee, and P. Berry. Interactive execution monitoring of agent teams. *JAIR*, 18:217–261, 2003.

[47] J. Yen, J. Yin, T. R. Ioerger, M. S. Miller, D. Xu, and R. A. Volz. CAST: Collaborative agents for simulating teamwork. In *IJCAI-01*, pages 1135–1144, 2001.