Kennedy, W.G., and Krueger, F. (2013) Towards Modeling Trust Behavior. Proceedings of the 12th International Conference on Cognitive Modeling (ICCM 2013). Ottawa. (11-14 July 2013)

Towards Modeling Trust Behavior

William G. Kennedy (wkennedy@gmu.edu)
Frank Krueger (fkrueger@gmu.edu)
Krasnow Institute for Advanced Study, 4400 University Drive
Fairfax, VA 22030 USA

Abstract

Previous work comparing fMRI data for two participants participating in a trust game provides a unique source for the development of an ACT-R model of trust. The model replicates a large portion of the behavior data. Continuing efforts expect to match more of the behavioral data and the imaging data comparison is expected to identify architectural needs.

Keywords: trust, cognitive modeling, ACT-R theory.

Introduction

Trust decisions are frequent cognitive activities and an integral part of social interactions, which seem to involve both rational and beyond rational cognition (Kennedy 2011, Kennedy et al., 2012). However, to develop an understanding of this mental process we have had only the observable outward behaviors, not the details on our internal mental processes. Brain imaging research has provided much more information about the internal processes. The project described here is a work toward developing a working theory of the mental process(s) of social trust using this new data.

The well-studied trust-related Prisoner's Dilemma game has players simultaneously decide whether to cooperate with their partner or to defect (Axelrod 1986). Since the simultaneity of the decisions complicates the processes involved, we used a similar but simpler version, a sequential two-person voluntary investment game (Berg et al., 1995). In this game, players alternated roles as the first or second decision makers (M1, M2). A players' decision tree is shown in Figure 1.

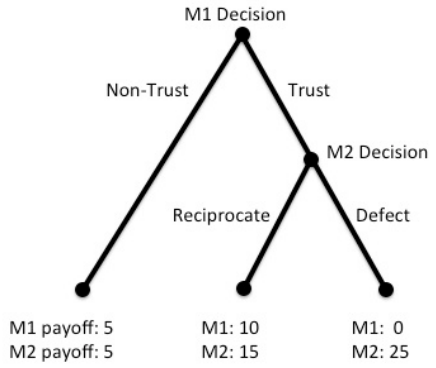

Figure 1. A Voluntary Trust Game Decision Tree

M1 decides to take a sure (and equal) payoff or to trust M2 and allow M2 to decide what payoffs each receives. The input to each player is the game's status. If M1 does not trust, M2 does not get to make a decision and both players receive an equal payoff. If M1 trusts, then the payoff is controlled by M2, who may reciprocate or defect. If M2 reciprocates, both players are better off in terms of their payoff. If M2 defects, then M2 gets a large payoff and M1 gets nothing. Note that in the this game, like the Prisoner's Dilemma game, mutual trust and cooperation lead to both players scoring well in the long term but not in the short term. Over the multiple runs with a pair of players, the tree layout is varied reversing the left-right sides of trust and non-trust options and reversing the sides of reciprocate and defect options. The payoffs are also varied in size, but the equality of the payoffs in the non-trust option and the ordering of the sizes of the payoffs were maintained.

The Human Data

Behavioral and imaging data was collected during a series of games played by 44 participants, organized as 11 pairs of women and 11 pairs of men (Krueger et al. 2007). The 22 pairs of strangers of the same gender played the game while being scanned in separate Magnetic Resonance Imaging (MRI) scanners. Each dyad played 36 games. Behavioral information (each decision) as well as the brain activation throughout the series of games was collected. The neuroscience side of this work is discussed elsewhere (Krueger et al. 2007).

The behavioral data shows that most pairs of players trusted each other. Four of the 22 pairs were perfectly consistent in their trusting and reciprocation exchanges. Ten of the 22 pairs had a few occasions of non-trusting or defections mixed in with predominately trusting and reciprocating exchanges. One dyad traded trusting and non-trusting exchanges throughout the 36 trials. There were four pairs that seem to change strategies over the trials (one from not trusting to trusting and three the other way). There were two dyads that had very few cooperative exchanges and one pair that did not complete the exercise.

There are several important characteristics of the observed behavior. First, for the vast majority of the interactions, both players were cooperative. Second, there were many examples of exchanges of uncooperative behavior, approximately balanced on both sides. Third, some human participants were repeatedly uncooperative

even when paired with very cooperative players. Finally, there were some indications that the varied reward payoffs did invite uncooperative behavior (Krueger, et al., 2007). This range of strategies is our challenge to model.

Computational Cognitive Modeling

The cognitive models of a player were developed using the ACT-R cognitive architecture (Anderson et al. 2004; Anderson 2007). The ACT-R system was interfaced with code that implemented the same protocol the human subjects experienced. ACT-R architecture generates a trace of its performance (at the millisecond level) as it processes sensory inputs, represents cognition, and generates output behavior. BOLD (blood oxygen level dependent) activation predictions applying fMRI are outputs also available from ACT-R. We have built models in ACT-R implementing theories of trust and replicating the human behavior data more and more.

The productions necessary to interact with the protocol environment and recognize the beginning of a series of games, the model's role in the next trial, read the displayed decision of the other player, read the payoffs shown on the tree diagram, and remember the history of the interactions and results require almost fifty productions. The heart of the mode is a much smaller set of productions that represent a theory of how humans make these decisions.

Modeling

We have built a series of models that are approaching matching the general human performance as well as its diversity. The first theory implemented was to maximize the player's payoff by evaluating what decision the second decider might make. We implemented a "like-me" Theory of Mind (Meltzoff 2007) in which the model evaluated what choice the other agent would make by placing itself in the role of the second decider and then using its own strategy. This decision is easy because the payoffs are known. However, this rarely matched human performance.

The second theory implemented was one of the two theories that were introduced in the original paper (Krueger et al. 2007), an "unconditional trust" strategy. As the first decider, the strategy was to always trusting and as second decider, to always reciprocate. This matches the vast majority of the human data (85.7% and 90.3% respectively).

The third strategy implemented was the "tit-for-tat" strategy (Axelrod 1986) and matched more of the interactions. When playing partners who were more frequently uncooperative, the model matches the other player's behavior, but does not match two characteristics of the exchanges. First, it does not initiate uncooperative play and second, it does not address the diversity in the play of dyads observed, i.e., some players continued to trust even when their partner was uncooperative.

Discussion

We are currently working on two more models. The first will implement a strategy that initiates less than always cooperative behavior by spontaneously. At issue is when to follow such a strategy. There is some evidence that the reward levels may tempt players to be uncooperative, but not all players. We are exploring the idea of individual differences, a threshold for boredom with a current strategy or curiosity or interest in exploring the effect of other options may explain the diversity observed. These ideas will be implemented in future models. When we have good correlation of our model with the previously collected behavioral data, we will compare ACT-R's imaging outputs with the previously collected data. This is expected to show active areas of the brain that are not represented by ACT-R's current mapping to the brain indicating.

Acknowledgments

This work was supported in part by AFOSR/AFRL grant FA9550-10-1-0385 and the George Mason University Center of Excellence in Neuroergonomics, Technology, and Cognition (CENTEC).

References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111*, 1036-1060.

Axelrod, R. (1986). *The Evolution of Cooperation.* New York, NY: Basic Books.

Berg, J., Dickhaut, J. and McCabe, K. (1995). Trust, Reciprocity and Social History. *Games and Economic Behavior, 10*, 122-42.

Kennedy, W. G. (2011). The Roots of Trust: Cognition Beyond Rational. *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society*, (pp.b188-193). Amsterdam: IOS Press.

Kennedy, W. G., Ritter, F. E., Lebiere, C., Juvina, I., Gratch, J. and Young, R. M. (2012) ICCM Symposium on Cognitive Modeling of Processes "Beyond Rational". *Proceedings of the 11th International Conference on Cognitive Modeling.* (pp. 55-58). Berlin.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, J., Zahn, R., Strenziok, M., Heinecke, A., and Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Science* 104(50): 20084-20089.

Meltzoff A. N. (2007). Imitation and Other Minds: The "Like Me" Hypothesis. In: S. Hurley and N. Chater (eds) Perspectives on Imitation: From Neuroscience to Social Science, pp 55-77. Cambridge: MIT Press.