Kennedy, W.G., and Krueger, F. (2013) Building a Cognitive Model of Trust Within ACT-R. Proceedings of the AAAI Spring Symposium 2013. Stanford, CA. (25-27 March 2013)

Building a Cognitive Model of Social Trust Within ACT-R

William G. Kennedy and Frank Krueger

Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA 22030 USA
wkennedy@gmu.edu, fkrueger@gmu.edu

Abstract

This paper describes work underway at the Krasnow Institute for Advanced Study on the topic of modeling social trust. We have built and are testing an ACT-R model intended to replicate human participants building and maintaining social trust using an economic investment game. We already have behavioral and fMRI imaging data for subjects which we expect to generate comparable data by having an ACT-R model read the same inputs the humans did and decide whether to trust or not their partner.

Introduction

Social trust is an everyday concept and tool that we all rely on, but may not understand well. It seems to be a judgment we make and then act based on that judgment. Its roots seem to be both rational and beyond rational (Kennedy 2011). 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. With brain imaging technology, we now have 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.

A way to test a theory is to compare the theory's predictions to data carefully collected from the actual phenomenon. (We may use some of the data to guide or make adjustments to a theory.) Economic games with repeated interactions in which trust is a component serve as a source of data and a test environment.

Probably the most famous trust-related game is the Prisoner's Dilemma in which players repeatedly decide whether to cooperate with their partner or to defect with non-zero sum payoffs (Axelrod 1984). In that game, the players make simultaneous decisions and only have the results of the previous interactions upon which to make their next decision. However, the Prisoner's Dilemma game is only a trust-related game in that the players seem to become focused on maximizing their score and miss the fact that if they both can be trusted to cooperate, they both maximize their score in the long run.

Here we use a version of the standard investment game developed by Berg et al. (1995). The players take turns in roles as the first or second decider and the second player knows the first player's decision before making his/her decision. In this way, trust is clearly the focus of the game. The first player makes a trust or not decision and the second player rewards that decision or takes one-sided advantage of that decision. In this experiment, pairs of strangers play many rounds of the game with the same person but with switching who makes the first decision. A players' decision tree is shown in Figure 1.

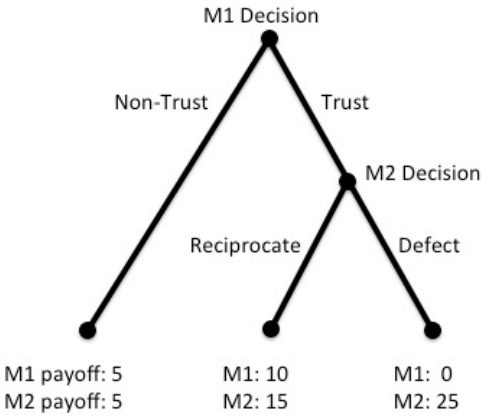

Figure 1. A Voluntary Trust Game Decision Tree

The first player decides whether to take a sure (and equal) payoff or to trust the other player, i.e., allow the other player to decide what both will receive as payoffs. In the display of the game's status, the selection is indicated by turning the line segment between the decision point, M1 or M2, downward to either a pair of payoffs or the M2

decision point. If the first player selects the non-trust branch (to the left in Figure 1 from the M1 Decision node), the second player does not get to make a decision. If the first player selects the trust branch (right side in Figure 1), then he or she turns over control of the payoffs to the second player. The second player then choses to either reciprocate or defect. If the second player reciprocates, both get payoffs. If the second player defects, that player gets a large payoff and the first, the originally trusting player gets nothing. Notice that 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 details of the protocol can be important. Figure 2 shows the sequence of images the two players see and the timing of the steps involved.

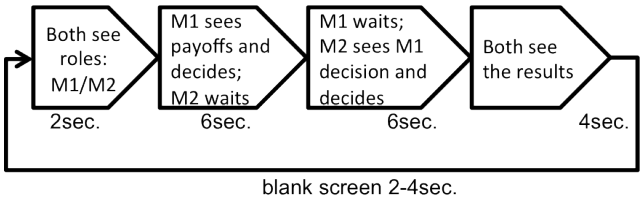

Figure 2. Voluntary Trust Game Protocol

The first step of the protocol is to establish the roles for the players, i.e., who makes the first decision and who makes the second. The second step in the protocol allows the first decider to see the decision tree with its payoffs and to decide whether to trust or to not trust. During this time, the second decider sees a blank screen. The displays are then reversed: the second decider sees the decision tree, payoffs, and the first player's decision. With that information, the second player decides whether to reciprocate or to defect. After the second player chooses left or right, both see the tree and the decisions made. If the first player decided not to trust, the second player's decision step is skipped and the system goes directly to showing the final results before showing a blank screen for a random time before starting the next game. These protocol details have been faithfully replicated in the cognitive model of the experiment.

The Human Data

We have data from 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 were observed in separate Magnetic Resonance Imaging (MRI) scanners and interacted through the voluntary trust game while being scanned. Each pair of partners played 36 games and 16 control games. The control games did not involve interactions between the players and each player simply made choices based on the payoffs available. The data collected includes the behavioral information on each decision as well as the brain activation throughout the series of games. For a discussion of the neuroscience side of this work, see (Krueger et al. 2007).

An overview of the behavioral data shows that most pairs of players trusted each other. Three 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. There was one pair where the players traded trusting and non-trusting exchanges throughout the 36 trials. There were four pairs that seem to switch strategies (one from not trusting to trusting and three moving the other way). There were two pairs that had very few trust-reciprocate exchanges and one pair that had none for half of the experiment and did not complete the other half. Table 1 shows this the counts of the different types of trials for each of the 22 pairs.

Table 1. Counts of Types of Exchanges in All Pairings (* marks pairing with 18 trials)

trust, reciprocate	trust, defect	non-trust
36	0	0
36	0	0
36	0	0
35	1	0
35	0	1
34	1	1
34	0	2
34	0	2
33	0	3
33	1	2
33	1	2
33	0	3
31	0	5
27	8	1
26	2	8
25	10	1
25	6	5
13	9	14
9	12	15
8	3	25
*1	*6	*11
0	27	9

In 19 of the 22 experimental parings, the first player's decision was to trust. However, in 3 of those cases the other player did not reciprocate. So, for 16 of the 22 first decisions was to trust and reciprocate. Overall, in 74.5% of the trials, the first player trusted and the second reciprocated. In the other cases, either the first player chose to not trust or the second player defected.

Table 2 shows the decisions for a pair who were in the group that seems to have switched strategies from generally trusting to non-trusting. In Table 2, the player who made the first decision is indicated by the recorded decision, whether the player chose to trust or not trust. The second player's decision is between reciprocating or defecting as shown in Figure 1.

Table 2. One Pair's Track Record

trial	player A	player B
1	trust	reciprocate
2	reciprocate	trust
3	trust	reciprocate
4	defect	trust
5	trust	defect
6		non-trust
7		non-trust
8	trust	reciprocate
9	defect	trust
10	non-trust	
11	reciprocate	trust
12	non-trust	
13	trust	defect
14	reciprocate	trust
15	trust	reciprocate
16	defect	trust
17	non-trust	
18	defect	trust
19	defect	trust
20	non-trust	
21		non-trust
22	trust	reciprocate
23		non-trust
24	non-trust	
25	non-trust	
26		non-trust
27	non-trust	
28		non-trust
29	trust	reciprocate
30	reciprocate	trust
31	reciprocate	trust
32	trust	reciprocate
33		non-trust
34	trust	reciprocate
35	defect	trust
36	trust	defect

It is interesting that in the data shown in Table 1, there were no long runs of players staying with a strategy, i.e., no trends. To test various theories explaining the participants' behavior, we developed and tested computational cognitive models.

Computational Cognitive Modeling

Computational cognitive modeling involves two steps: code to implement the experiment's protocol and the running of the cognitive model. We have developed the code to provide the task's process and we have developed a series of computational cognitive models of a human's perception of the task, decision-making, and execution of those decisions. The protocol code presents the images associated with the protocol to the users, which may be either humans, cognitive models, data from previous experiments, or any combination of two. For each trial, the system starts by presenting the assigned roles of the players. This code then manages the execution of the rest of the protocol shown in Figure 2 for both players.

The cognitive models of a player were developed using the ACT-R cognitive architecture (Anderson et al. 2004; Anderson 2007). A cognitive architecture is defined as:

> A *cognitive architecture* is a specification of the structure of the brain at a level of abstraction that explains how it achieves the function of the mind. (Anderson 2007, pg 7)

The ACT-R system interfaces with the protocol system and supports perception, reasoning, and implementation of decisions.

Computational cognitive modeling is the implementation of a theory of cognition that produces predictions of human cognition. A model is "software" in the form of decision rules and the prior knowledge that runs on the "hardware" of a cognitive architecture, which provides the cognitive functions. ACT-R is an example of an architecture (Anderson et al. 2004; Anderson 2007) and it provides a trace of its performance at the millisecond level as it processes sensory inputs, performs decision-making, and generates output behavior. It can also produce BOLD (blood oxygen level dependent) activation predictions applying fMRI. We have implemented this trust game in a form that an ACT-R model can interact with it directly, i.e., the model can "see" the display and make decisions.

Our cognitive models 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. These functions alone require almost fifty productions. The core of the cognitive model is the set of productions that decide whether to trust or not when the model is the first decision maker and whether to reciprocate or defect when the second player. These productions implement a theory of how humans make these decisions.

One of the "if-then" productions associated with deciding which option to take at the M1 node is shown below and will be explained line by line.

```
(P M1-prepare-normal-right
   =goal>
      isa         game
      own-role    m1
      dec1        ready
      Rpayoff     nil
    =imaginal>
      isa         payoffs
      type        normal
      p6          =x
==>
    =goal>
      Rpayoff     =x
    =imaginal> )
```

The first line starts the production and contains the name of the production. This production will add to the goal buffer the value of the payoff if the M1 decision were to select the right option when the decision tree is of type "normal". This will be later used to decide whether to select this option. The second line begins the specification of the "if" part of the production. In this production, the "if" part involves the goal buffer and the "imaginal" buffer. Both buffers contain slots and their required value for this production to fire resulting in the implement the "then" part. The "isa" slot specifies what type of knowledge chunk is in the buffer. In this case, the type of chunk in the goal buffer must be a "game". The next slot identifies that the model's role in this trial must be to go first and make the decision at the M1 node. The "dec1" slot has is expected to have the value "ready" indicating that the system is working on the M1 decision. Finally, the next line requires that the payoff to the right be not been filled, i.e., is nil. The second buffer tested in the "if" part is the "imaginal" buffer. This buffer is of type "payoffs" indicating that it holds the payoffs read in previously. The "type" slot check that this rule will only apply to cases when the payoff tree type is "normal", meaning the non-trust option is to the right. The payoff value for going to the right for a normal tree is payoff number 6 as read in. The line "p6 =x" collects the value of that slot in a variable "x" for later use. The symbol "==>" indicates the transition of the production to the "then" part. If the conditions are met, the "then" part indicates that the payoff value for going to the right in the goal buffer will be set equal to the one that was in the "p6" slot of the imaginal buffer. The last line of the production references the imaginal buffer on the "then" side of the production so that the ACT-R system will retain the buffer's contents. If a buffer is referenced on the "if" side but not on the "then" side, the system would automatically clear that buffer, which is not desired by this production.

This production fills in the easily identified payoff for deciding to select the right option. There is another production very much like this one for the opposite decision tree layout. Other productions implement the Theory of Mind evaluation of what the payoff would be for selecting the other option. Then, when both the payoffs to the left and to the right are known, another production would fire and select the option with the larger payoff. This is only one possible theory of how people decide how make their decision.

The first theory that was implemented was to pick the choice that would maximize the player's payoff. As the first decider, this requires 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 approach. This decision is easy because the payoffs are known. Then, with the model returns to the role of the first decider and uses the results of the simulation of the other, to provide the payoff for both options and it can make its decision.

This model was relatively straightforward to implement, taking only 15 rules, but for this strategy, the results are that it consistently selected the non-trust option when the first decider and the defect option when the second decider. In addition, the behavior does not match or relate hardly at all to the observed behavior in the human data.

The second theory that was implemented was one of the two theories that were were introduced in the original paper (Krueger et al. 2007), an "unconditional trust" strategy. In the reference, this theory was described as a learned behavior based on the lower response times, which seemed to indicate the evaluation process was reduced to a simple heuristic. The learned strategy of the first decider always trusting and the second decider always reciprocating does match much of the human data, matching the initial behavior of most teams (16 of 22 first decisions) and 74.5% of all trials.

Discussion

While matching almost 75% of the experimental data is easily possible, it is not satisfying. There is obviously

interesting behavior to be modeled in our cognitive model. The next cognitive model is intended to replicate the occasional uncooperative behavior of not trusting the other player or the more confrontational defecting behavior. Although this could be done randomly, it would not be cognitively plausible.

Table 3 shows the decisions for a pair who were in the group that seems to have switched strategies from generally trusting to non-trusting.

Table 3. One Pair's Track Record

trial	player A	player B
1	trust	defect
2	reciprocate	trust
3	trust	reciprocate
4	reciprocate	trust
5	trust	defect
6	defect	trust
7	reciprocate	trust
8	trust	reciprocate
9	reciprocate	trust
10	trust	defect
11	defect	trust
12	trust	defect
13	trust	defect
14	defect	trust
15	trust	defect
16		non-trust
17	trust	defect
18	defect	trust
19	defect	trust
20	non-trust	
21	defect	trust
22	non-trust	
23		non-trust
24	non-trust	
25	non-trust	
26	reciprocate	trust
27	trust	reciprocate
28	reciprocate	trust
29	trust	defect
30		non-trust
31	defect	trust
32	non-trust	
33		non-trust
34	non-trust	
35		non-trust
36	non-trust	

In Table 3, the first decision is indicated by whether the player chose to trust or not trust and the second player's decision is between reciprocating or defecting as shown in Figure 1. In the first 15 trials, the first decider always trusted and the second decider sometimes reciprocated and sometimes defected. It was not until the 16th trial that a first decider executed the alternative to trusting, non-trusting, and then it was several more trials before they both exercised that option. By the end of the set, the two players went five consecutive trials of no trusting choices.

There appears to be evidence of some exploring of options, i.e., to try another decision option and see what happens after 15 trials. Something similar appeared in the data for 10 or more pairings as shown in Table 1 in which there are only a small number of trials that were not trust-reciprocate.

Studying the evidence further, we could interpret the result as the "tit for tat" behavior that won the Prisoner's Dilemma competition (Axelrod 1986). Axelrod ran a competition of submitted strategies implemented in computer programs and paired the submissions. The top scoring program began cooperating but if the other agent defected, it did so on the next turn. This behavior turned out to be the highest scoring against the other implemented strategies. So, our next cognitive model should include the exploring behavior and a "tit-for-tat" strategy.

Following that cognitive model, we plan to develop a model that learns from the exchanges and adjusts its behavior. We believe that will replicate the observed behavior where players appear to be changing their strategy over time. There are more clear examples in the data, but Table 3 shows that the exploration led to more exploring, i.e., trying the defect option, and ended up with both players selecting the non-trusting option for the last five runs.

In addition to improving the sophistication of our cognitive model, we also have the response time data for the human participants. The ACT-R architecture generates this output for runs if models and adjusting our model to produce comparable response times improves the credibility of our model.

Finally, the original focus of the neurological experiments was to study the regions of the brain involved with trust. The ACT-R architecture is capable of producing similar data. If our cognitive model can also produce neuroimaging data that compares well with the human data, that further enhances the credibility of our cognitive model. We anticipate some difference between our model's neuroimaging data and the human data because the ACT-R architecture does not (yet) address the ability of humans to reason about the reasoning of others, i.e., Theory of Mind. Therefore, we expect that our model will not match the human imaging data associated with that function.

A computational cognitive model that implements a potential theory of trust for the volunteer trust game can, through the ACT-R system, produce behavioral and imaging data that can be compared to the human participants' data. This includes replicating the behavioral

data overall, response time data, and the rate of learning/change in behavior with experience. The human participants' brain activation can also be compared to the ACT-R model's output. Through these comparisons, we will evaluate the goodness of fit (Schunn and Wallach 2005) for the implemented theories and support claims of the models representation of the humans' cognitive processes of trust (Fum, Del Missier, and Stocco 2007).

Conclusions

We have developed a computational model that uses a Theory of Mind simulation of another agent to decide whether to trust or not. However, a simple trusting model would perform better, at the level of matching almost 75% of the human trial data. We have also demonstrated that having human data is very useful for developing theories of trust behavior and that cognitive modeling is a useful approach to testing theories of cognition. At this stage of our research, we are attempting to match the overall behavior, it terms of the number of times a player makes different decisions over the set of 36 trials. We have not yet attempted to match trends in behavior, individual decisions, or response times. Those measures will be addressed in further research.

References

Anderson, J. R. 2007. *How Can the Human Mind Occur in the Physical Universe?* New York, NY: 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(4): 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.

Fum, D., Del Missier, F., and Stocco, A. 2007. The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research* 8:135-142.

Kennedy, W. G. 2011. The Roots of Trust: Cognition Beyond Rational. In *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society*, 188-193. Amsterdam: IOS Press.

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.

Schunn, C. D., and Wallach, D. 2005. Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), Psychologie der Kognition: Reden and Vortra¨ge anla¨sslich der Emeritierung von Werner Tack: 115–154. Saarbruecken, Germany: University of Saarland Press.